Segmentation of Oral Optical Coherence Tomography with Deep Learning

by

Chloe Danica Hill

BASc, Simon Fraser University, 2020

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Applied Science

in the
School of Engineering Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Summer 2023

Declaration of Committee

Name:	**Chloe Danica Hill**
Degree:	**Master of Applied Science**
Title:	**Segmentation of Oral Optical Coherence Tomography with Deep Learning**
Committee:	**Chair: Shawn Sederberg** Assistant Professor, Engineering Science
	Pierre Lane Supervisor Associate Professor, Engineering Science
	Mirza Faisal Beg Committee Member Professor, Engineering Science
	Andrew Rawicz Committee Member Professor, Engineering Science
	Ivan Bajić Examiner Professor, Engineering Science

Abstract

Diagnosis of oral cancer involves collecting multiple biopsies to increase the likelihood of sampling the most pathologic site within a lesion. Optical coherence tomography (OCT) allows for examination of subsurface morphology, and has shown potential in biopsy guidance. OCT captures changes in tissue stratification related to depth, topology, and presence of the stromal-epithelial boundary which are structural biomarkers for pre-invasive and invasive oral cancer. This thesis presents a four-part neural network pipeline to simplify OCT interpretation by providing *en face* maps of epithelial depth and stratification. U-nets models are employed to segment the stromal-epithelial boundary, and supporting convolutional neural networks are used for identification of the imaging field and artifacts. Training was conducted on a variety of non-cancerous and cancerous pathologies across the oral cavity to promote generalizability. Predictions demonstrate as-good-as or better agreement than inter-rater agreement, suggesting strong predictive power.

Keywords: optical coherence tomography; oral cancer; cancer imaging; *in vivo*; deep learning; image segmentation

Acknowledgements

Foremost, I wish to express my deepest gratitude to my supervisor, Dr. Pierre Lane, for his unwavering support and dedication throughout my research journey. His belief in my abilities and his expert guidance has been instrumental in shaping my work. The supportive and curiosity-driven atmosphere he cultivates in his lab added much to my graduate experience and pushed me explore new ideas and perspectives. I would also like to recognize Dr. Andrew Rawicz and Dr. Mirza Faisal Beg for their expertise, understanding and patience throughout my time at SFU. Their commitment to my research and academic success has contributed significantly to the quality and depth of my work.

I am very grateful to have been welcomed into the Optical Cancer Imaging Lab (OCIL) at the BC Cancer Research Institute. I would especially like to recognize Dr. Calum MacAulay for his expertise and many contributions throughout my research. Additionally, I would like to recognize Dr. Catherine Poh, whose work provided the foundation for this research. Finally, I would like to recognize my fellow students, Ms. Jeanie Malone, Mr. Eric Brace, Mr. Adrian Tanskanen and Mr. Ian Janzen. The members of the OCIL lab have been a constant source of support, guidance and encouragement. Their willingness to share their expertise and provide feedback as well as ensure polished presentations has been invaluable in refining my research. The friendships and shared passions fostered in this lab have made this journey both enriching and enjoyable.

There are more people who have played a role in the realization of this thesis than I can mention, but I would to thank all of my friends and family who have supported me throughout these years. I would not be here without you.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AUC	Area Under the Curve
BCE	Binary Cross Entropy
CIS	Carcinoma *In Situ*
CNN	Convolutional Neural Network
DL	Deep Learning
DSC	Dice Similarity Coefficient
FN	False Negative
FOV	Field of View
FP	False Positive
LR	Learning Rate
mAP	Mean Average Precision
NA	Negative Agreement
NN	Neural Network
OCIL	Optical Cancer Imaging Lab
OCT	Optical Coherence Tomography
OED	Oral Epithelial Dysplasia
OPMD	Oral Potentially Malignant Disorder
OSCC	Oral Squamous Cell Carcinoma
PA	Positive Agreement
Po	Proportion of Observed Agreement
PRC	Precision-Recall
ROC	Regional Operating Curve
RPD	Rotary Pullback Drive
SD	Standard Deviation
SFU	Simon Fraser University
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate

VC	Verrucous Carcinoma

Chapter 1. Introduction

1.1. Motivation & Objectives

Early detection and diagnosis of cancer improves patient prognosis and potential for successful treatment. In head and neck cancers, the 5-year survival rate is 86% for localized cancers, but decreases to 69% for regional cancers and 40% for distant metastatic cancers. Unfortunately, only 28% of head an neck cancers are detected at the localized stage [1]. Screening methods consist primarily of incisional biopsy and histopathologic examination, which is a burden on both the patient and the healthcare system. The most common treatment for oral cancer is surgical resection; this procedure can result in devastating physiological and psychological effects [2].

The utility of optical coherence tomography (OCT) as an adjunct screening tool for oral cancer has been previously demonstrated [3]–[6]. OCT allows for minimally invasive volumetric examination of subsurface tissue structures. Notably, thickening and changes in stratification of the epithelial layer are biomarkers for oral cancer, which can be visualized through OCT. Yet, clinical adoption of this technology requires data analysis tools to provide rapid and reproducible assessment of tissue morphology, as the large volume of data collected during imaging makes manual assessment intractable. A possible solution is deep learning (DL) algorithms which are well suited to processing large amounts of data. DL methods are effective for bulky datasets, requiring minimal interference once trained, and generate efficient and consistent outcomes over high volumes of repetitive tasks. DL methods have been previously applied to generate segmentation of OCT, primarily in ophthalmologic OCT data [7]–[9], but also in esophageal [10], and intravascular OCT data [11].

In this thesis, I explore using DL segmentation techniques to (1) detect the presence and (2) quantify the thickness of the epithelial layer in oral OCT. A retrospective study is conducted to assess whether neural net can provide repeatable segmentations with comparable accuracy to manual annotation, at near-real time. Clinical applications of this work could include integration into the diagnostic workflow for incisional biopsy guidance and identification of tumour margins during surgical excision.

1.2. Previous & Related Work

The junction of deep learning and OCT image analysis has been well documented, with substantial evidence indicating utility in image segmentation tasks [8], [12]–[15] and classification tasks [16]–[18]. The bulk of research has been applied to ophthalmologic OCT, where OCT is has become an ancillary step for monitoring and diagnosis of several eye disorders and conditions, such as age-related macular degeneration [19]. Comparatively, image analysis of oral OCT with DL is an emerging field with relatively few publications. Recent work has explored applying classical [20], out-of-the-box DL [21], fine-tuned DL [22], and machine learning combined with DL [22] methods to triage healthy, pre cancerous and cancerous OCT image samples, and to classify states of dysplasia [21].

The BC Cancer Optical Cancer Imaging Lab (OCIL) previously developed segmentation software [23] for oral OCT, using classical image processing techniques to identify the top and bottom surfaces of the epithelial layer. The work presented in this thesis follows similar methodology but aims to overcome inherent limitations of classical segmentation approaches. Specifically, this previous approach required manual per-image parameter adjustments and took up to 60 seconds to process each image, despite software optimizations.

While previous attempts have demonstrated moderate success, classical methods suffer from poor generalizability, extensive processing times and high computational demands, resulting in poor clinical applicability. Conversely, previous DL approaches have excelled in fast processing times and reduced computation demands, but prior applications have focused on classification tasks, aiming to provide diagnostic information about tissue status. There is a gap in the existing literature for image analysis of oral OCT, where no pathology agnostic, site agnostic, rapid and repeatable tool exists to identify structures of interest. We propose that this tool should not provide diagnostic suggestions, but instead empower clinicians and clinical decision-making by providing additional data through easily interpreted visualizations of subsurface morphology.

The work presented this thesis is an extension of research performed during my undergraduate thesis [24]. My previous research looked to solve the same goal of oral mucosa epithelial layer detection, through application of deep learning techniques. Moderate success in the initial approach provided a rudimentary proof of concept that this task was achievable, and elucidated gaps in both dataset and methodology. This work provided a fundamental platform to broaden the research scope and build a more generalizable and comprehensive approach.

1.3. Contributions

DL algorithms for OCT interpretation should be auxiliary to the clinical workflow, providing clinicians with the necessary tools to utilize complex imaging techniques, without providing diagnostic information. This work presents a unique approach to the analysis of oral cancer, aiming to detect tissue features without providing information on the health of the sample. To my knowledge, it is the first approach applying DL segmentation methods to oral OCT. Without the constraints of rigid classification labels, I was able to develop a more generalized approach, confined to the oral cavity, but not limited to specific oral sites or diseases. Rather, the goal of this work was to be invariant to tissue type and pathology, and instead provide reproducible identification of the epithelial layer boundaries. I presented this work at the SPIE Photonics West Conferences in 2022 (Multimodal Biomedical Imaging XVII) and 2023 (Imaging, Therapeutics, and Advanced Technology in Head and Neck Surgery and Otolaryngology) [25], [26]. An additional outcome of this work was the development of a MATLAB application, designed specifically for manual segmentation of OCT. This application has already been utilized to develop annotation data for pulmonary, gynecological and *en face* oral OCT data.

1.4. Chapter Organization

Chapter 2 of this document presents relevant background information, including oral physiology and carcinogenesis (Section 2.1), OCT hardware, acquisition and geometry, (Section 2.2), DL methods and applicable terminology (Section 2.3), and the metrics used to evaluate DL models (Section 2.4).

Chapter 3 describes the methods used to create the reference data set (Sections 3.1 and 3.2), including the development of a MATLAB application (Section 3.3), data selection and rater training (Section 3.4). Chapter 4 covers the implementation and results of four independent neural networks to generate near real time: (1) location of the imaging field of view (FOV; Section 4.2), (2) identification of the epithelial surface layer (Section 4.3), (3) identification of location and discontinuities in the basement membrane (Section 4.4), and (4) detection and isolation of image artifacts due to air bubbles or markers within the imaging sheath (Section 4.5). This document discusses future directions and concludes in Chapter 5.

Chapter 2. Background

2.1. Physiology of the Oral Cavity

The OCT and biopsies described in this thesis were collected from various sites within the oral cavity. Summarized in Figure 2.1, sites include: the tongue (ventral, lateral and dorsal), the gingiva or gum, the floor of the mouth, the roof of the mouth (including the hard and soft palate), the cheek lining (which is comprised of the buccal mucosa), the labial mucosa, the vestibule, and the lips. Discussed below, specific tissue composition varies across sites, resulting in changes to mechanical and optical properties.

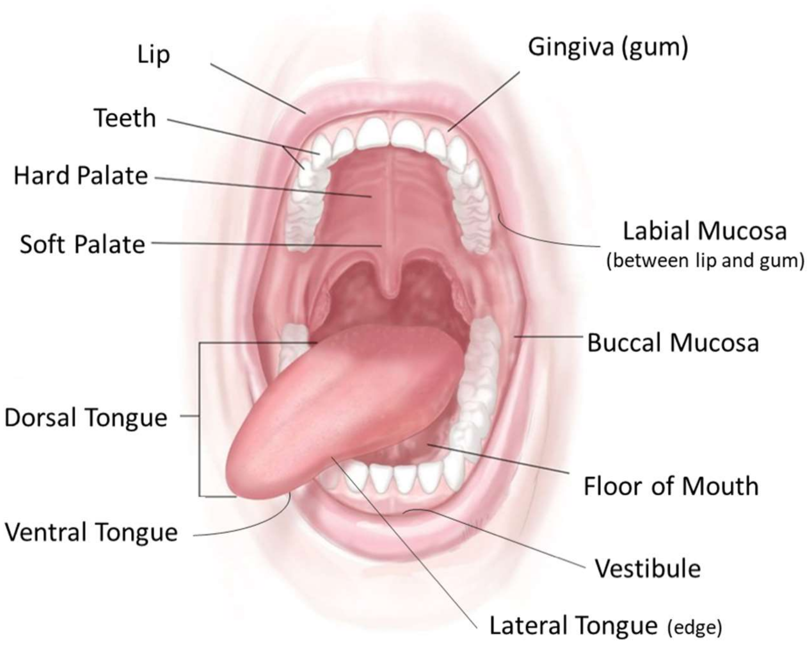


Figure 2.1. Anatomical sites of the oral cavity.
Source: Adapted from figure for the National Cancer Institute © 2012 Terese Winslow LLC, U.S. Govt. has certain rights [27].

Shown in Figure 2.2, the oral mucosa typically consists of three layers: the *epithelial layer* composed of avascular, stratified squamous epithelial cells; the *stromal* or *lamina propria* layer, a connective tissue layer composed of blood vessels and nerves among others; and the *submucosa*, an occasional layer of fibrocollagenous and elastic tissue. The epithelial and stromal layers are separated by a thin non-cellular basement membrane [28]. Depending on location, the oral mucosa may be *keratinized*, *para-keratinized* or *non-keratinized,* which describes unique organizations of the epithelial layer.

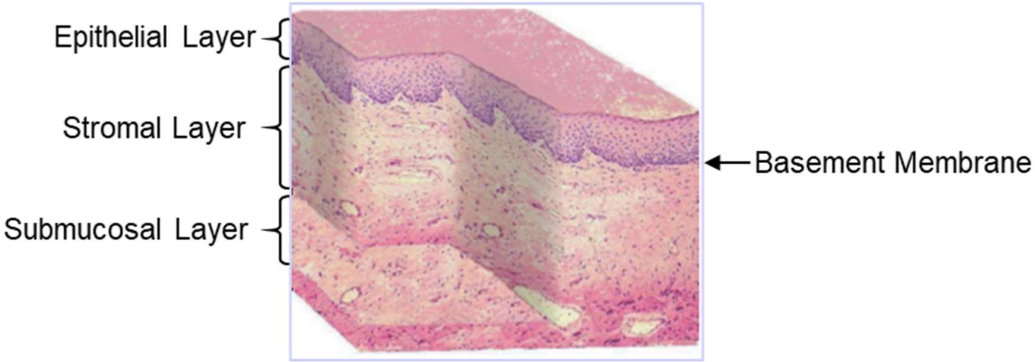


Figure 2.2. Layered morphology of normal oral mucosa.
Adapted from [29].

Histological, functional, clinical, and site differences classify the mucosa into distinct phenotypes: *lining mucosa*, which is non-keratinized and present in mobile regions; *masticatory mucosa*, which contains no submucosal layer and may be keratinized or para-keratinized to provide support for structured regions; or *specialized mucosa*, which contains nerve endings for sensory and taste perception, may be keratinized or non-keratinized. Lining mucosa covers soft or mobile regions such as the soft palate, ventral tongue, floor of mouth, lips, labial and buccal mucosa. The epithelial layer of the lining mucosa is typically thin (80-200 µm) [30]–[32], excluding labial and buccal mucosae, which present thick epithelium (300-500µm) [30], [32], [33] The gingiva and hard palate are covered by masticatory mucosa, while the dorsal and lateral tongue are composed from specialized mucosa. These regions are subjected to higher forces during mastication, and exhibit moderately thick epithelial layers (250µm) [30].

Layers of the oral mucosa layers can be visualized with histological staining of biopsied tissue. The preferred stain in cancer diagnosis is hematoxylin and eosin (H&E) stain, which allows for identification of nuclei (purple/blue) and cytoplasm and extracellular matrix (pink) [34]. Histology of the specialized mucosa of the lateral tongue is shown and annotated in Figure 2.3.

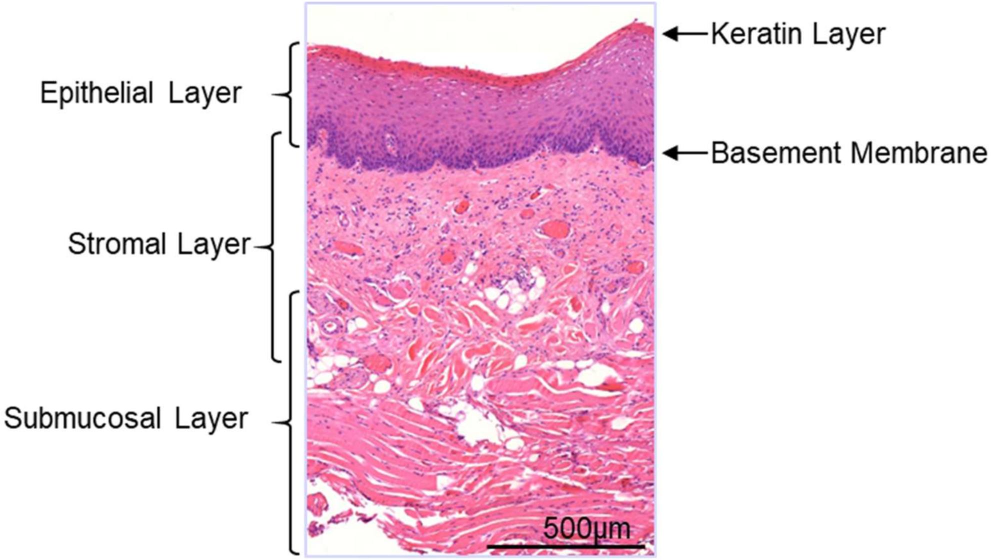


Figure 2.3. **Annotated H&E histology of normal tissue from lateral tongue (specialized mucosa).**

2.1.1. Screening & Diagnosis of Oral Cancer

It was estimated that 7,500 Canadians were diagnosed with head and neck cancers in 2022, and 2,100 succumbed [35]. The prognosis of oral cancer worsens dramatically with late-stage detection; when detected at early stages (local cancers), the 5-year survival rate is 85%, decreasing to 68% and 40% in loco-regional and metastatic stages, respectively [1]. Current screening methods consist of white light exploration, alongside diagnostic adjuncts such as toluidine blue staining and auto-fluorescence imaging [36]. These techniques are limited to surface examination and provide no information on the status of the basement membrane.

Suspect lesions require biopsy and histopathological assessment for definitive diagnosis. Further discussed in Section 2.1.2, there are several structural biomarkers for invasive cancer visualized through histological staining, but ensuring that the sample contains the most pathologic tissue is difficult. Shown in Figure 2.4, clinical presentation benign lesions (panel a) may appear similar to occult lesions (panel b), multiple biopsies may be taken to prevent false negatives [37].

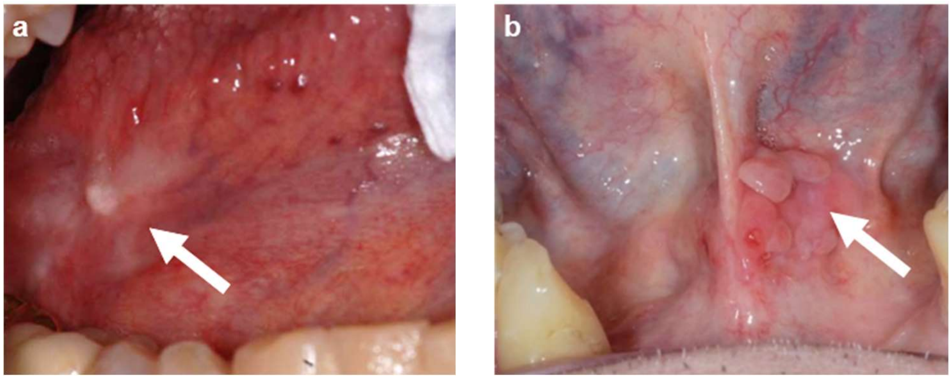


Figure 2.4. Comparison of white light imaging of suspicious lesions in the oral cavity: (a) benign lesion on ventral tongue; (b) cancerous lesion on ventral tongue. White arrows denote suspicious lesions.

2.1.2. Progression of Oral Cancer

The multi-step progression from healthy to cancerous oral tissue is a well documented phenomenon. Dysregulation typically originates at the basement membrane and is not visually evident at the tissue surface during initial stages. Figure 2.5 summarizes the progression of oral cancer at the cellular level. Lesions arising from changes to cell shape or size, or increase in proliferation are classified as benign, potentially malignant or malignant [37].

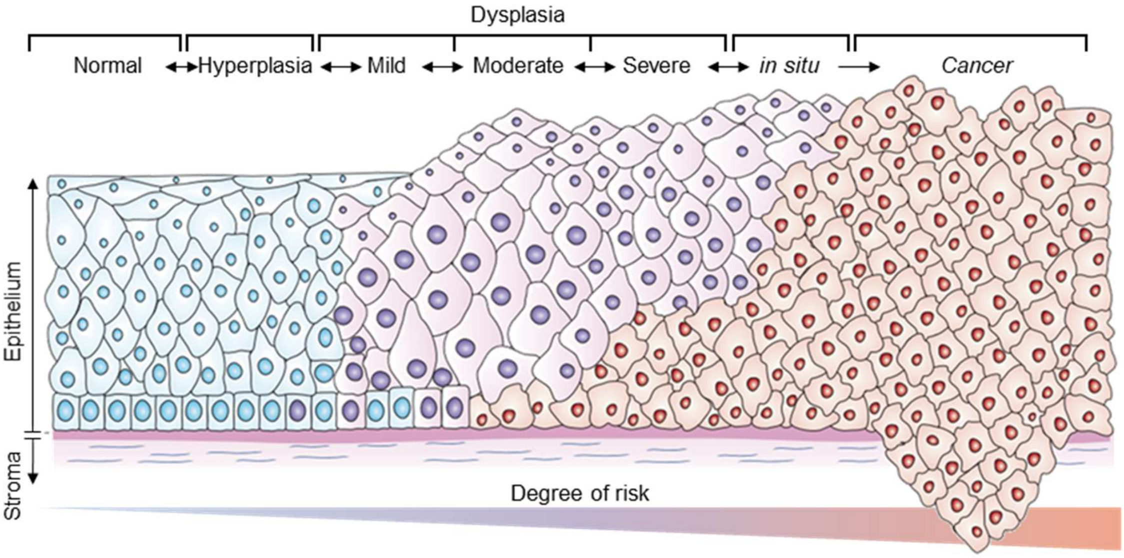


Figure 2.5. Cellular progression of oral tissue from normal morphology to cancerous.
Adapted from [38].

Benign oral lesions are non-cancerous growths that include fungal infections, frictional lesions and hyperplastic lesions [37], [39], among others. In this research, benign lesions are limited to melanotic macules (small, pigmented regions due to an increase in melanin), candidiasis (fungal infection), hyperkeratosis (abnormal thickening of keratin layer), reactive hyperplasia (excess cellular proliferation arising from trauma) and scars.

Oral epithelial dysplasia (OED) is used to qualify the histomorphologic presentation of pre- or potentially malignant lesions. Analysis of histological features is used in complement with clinical presentation when diagnosing oral potentially malignant disorders (OPMDs) [40], [41]. The World Health Organization has identified architectural (presence and degree of epithelial stratification), and cytological (cellular atypia) changes as key indicators of dysplastic progression [42]. A 4-tier grading system of mild (grade 1), moderate (grade 2), severe (grade 3) dysplasia, and *in situ* neoplasm is used to define the quantify OED. *In situ* lesions are sometimes synonymous with severe dysplasia. This research includes examples of OPMDs (verrucous hyperplasia and actinic cheilitis), as well as OED grades 1 through 3, carcinoma *in situ* (CIS) and lentigo maligna (*in situ* melanoma).

Malignant lesions are classified by having breached the basement membrane, breaking the barrier preventing the spread of cancerous cells into connective tissue, and allowing for potential metastasis [43]. The most prevalent invasive tumour is oral squamous cell carcinoma (OSCC), accounting for over 90% of oral tumours [44]. Other cancers include verrucous carcinoma (VC; a subset of OSCC) and lentigo maligna melanoma.

2.2. Optical Coherence Tomography (OCT)

Optical coherence tomography (OCT) is an optical imaging technique that generates volumetric data through the reconstruction of backscattered signal generated from a low coherence light source [45]. OCT bridges the resolution-gap between ultrasound and microscopy, imaging at a resolution of 1-15μm and providing information at the microstructure level which was previously only available through biopsy. OCT is sometimes referred to as an optical biopsy, as the morphological features imaged are strongly correlated to those observed in histology [45]. OCT is most commonly used in ophthalmology [46], [47], but has applications in a variety of other fields, including dermatology [48], [49], cardiology [50], pulmonology [51], gynaecology [52], and dentistry [53]. OCT collects real time images with minimally invasive, non-destructive methods, without the use of ionizing radiation or contrast mediums. Moreover, low-cost components and system portability allow for use in surgical suites [54].

A consequence of optical imaging is an inherent trade-off between imaging depth and resolution. This, in addition to the rapid attenuation of visible/near infrared light [54] into tissue means that the core limitation of OCT is a shallow imaging depth, collecting data at no more than 2-3mm in depth. However, the depth achieved by OCT is comparable to the depth of histological data collection, making it an excellent tool for examining changes to near-surface tissue [55].

2.2.1. System Hardware

The system used to collect data for this study has been detailed in a previous publication [6]. In summary, a 1310 ± 50nm polarization swept source OCT system and rotary pullback drive (RPD) was developed to capture wide-field images in the oral cavity, shown in Figure 2.6, panel a. The primary advantage of this system is an ability to collect images up to 90mm in length, allowing large tissue sites to be collected in a single scan (referred to as a pullback). Two catheter sheath holders were developed to allow for imaging of various sites in the oral cavity: a modified dental mirror (panel b, top) and a modified saliva injector (panel b, bottom). A single mode fiber serves as the light delivery and signal collection system. During *in vivo* imaging, an optical fiber is packaged to enable rotational scanning (panel c).

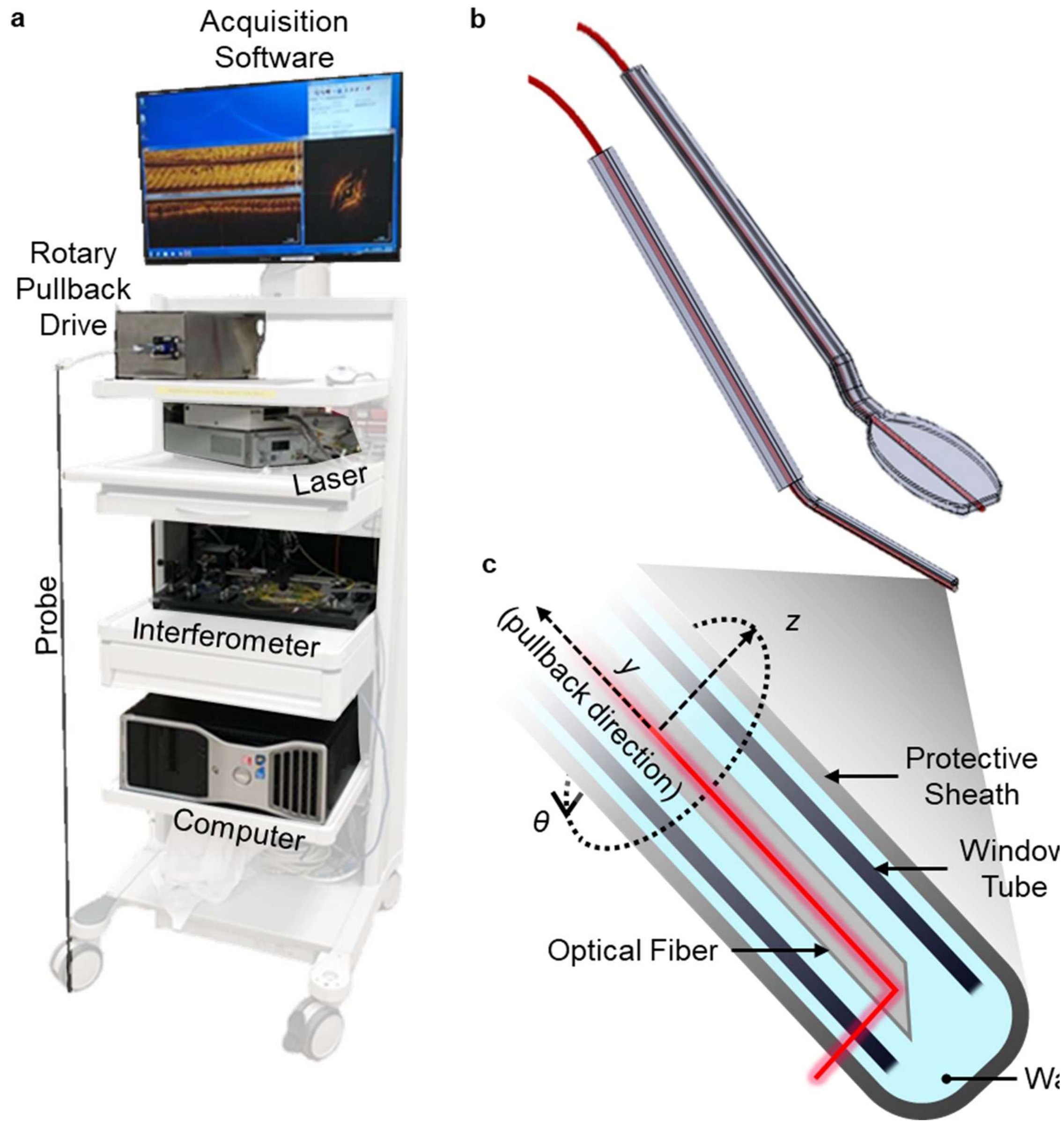


Figure 2.6. OCT Collection hardware; (a) clincal imaging tower; (b) catheter holders for fiber-optic oral OCT; (c) optical fiber wrapped in window tube with helical scanning coordinates.

The two catheter holders were developed to allow versatile imaging, with the dental mirror probe providing improved imaging of planar surfaces (e.g. tongue and buccal mucosa), and the flexible modified saliva injector probe allowing easier placement in restrictive sites (e.g. floor of mouth, gingiva). Representative clinical photos for various sites are shown in Figure 2.7, with examples of a lateral tongue imaging being collected with a modified saliva injector probe (panel a), floor of mouth imaging being collected with a modified saliva injector probe (panel b), and ventral tongue imaging being collected with the modified dental mirror probe (panel c).

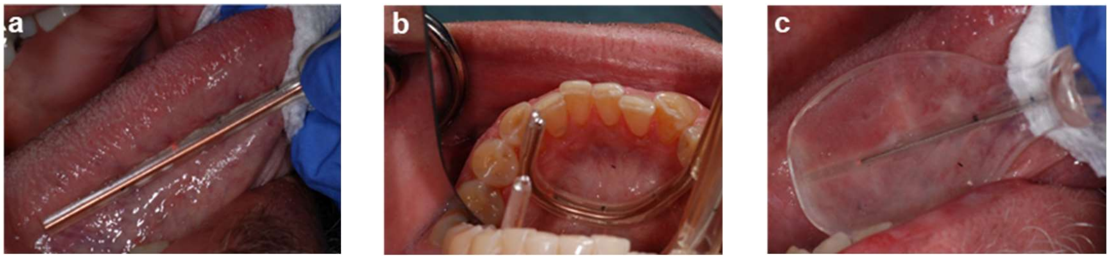


Figure 2.7. Clinical photos of (a) lateral tongue (b) floor of mouth being imaged with modified saliva injector probe; (c) ventral tongue being imaged with modified dental mirror probe.

2.2.2. OCT Orientation

Widefield OCT data is acquired in a cylindrical 3-dimensional volume. The coordinate system defined in Figure 2.8 (panel a) includes radial or a-line axis (*z*), azimuthal (*θ*), and pullback (*y*) axes. The same coordinate system can be seen overlaid on the imaging probe above, in Figure 2.6 (panel c). Slicing the volume perpendicular to the pullback axis allows for visualization of 2-dimensional cross-sectional views (cross-sectional b-frames), which can be viewed in collected in polar coordinates (panel b), or unwrapped into Cartesian coordinates (panel c). Unwrapping the volume along the azimuthal axis and applying a mean intensity projection along the radial axis, creates a 2-dimensional *en face* image (panel d). Slicing the volume perpendicular to the azimuthal axis allows for visualization of the longitudinal view (panel e, inset panel f). This work is conducted primarily on the longitudinal view, as it encompasses the most information and presents the most similar to histology.

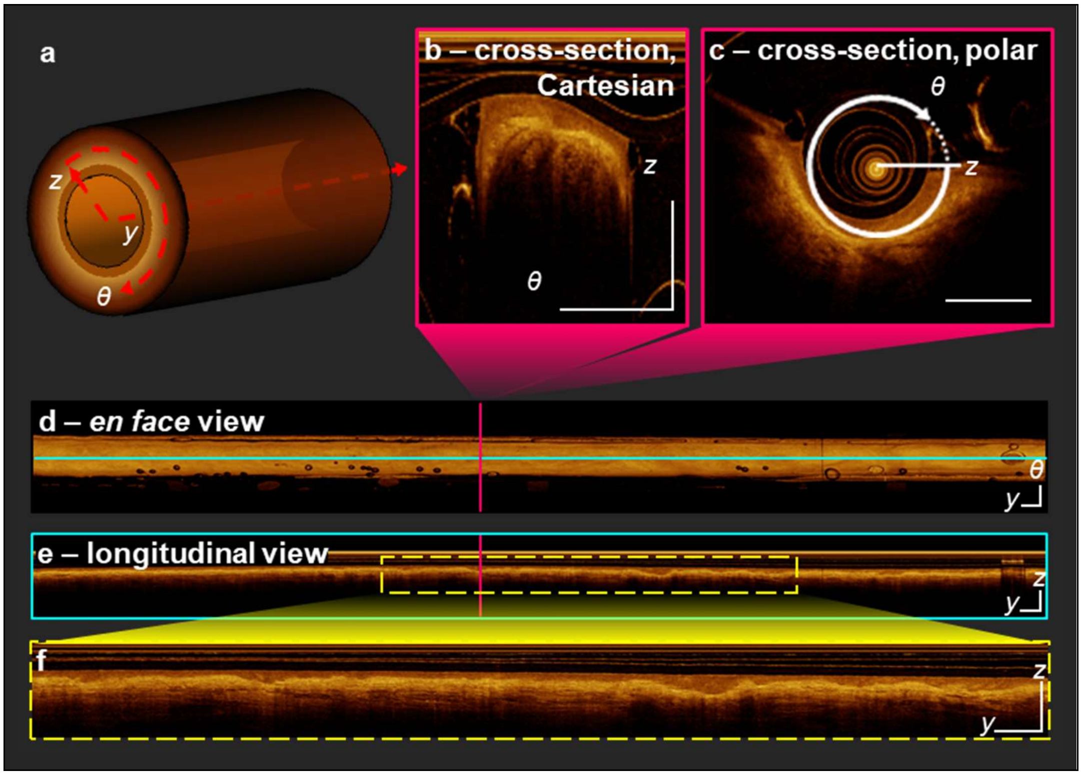


Figure 2.8. **Orientation of OCT pullbacks. (a) Coordinate system; (b) Unwrapped and (c) wrapped cross-sectional view of a single slice captured at the vertical pink line shown in (d) en face projection; (e) Longitudinal view of a single slice captured at the horizontal cyan line in the en face projection; (f) Insert view of yellow dashed box, detailing information captured in the longitudinal view. Scale bars 1mm.**

2.2.3. Data Collection

In vivo imaging of the oral mucosa using the system described Section 2.2.1 was approved by the Research Ethics Board of the University of British Columbia and the British Columbia Cancer Agency (H11-02516). Images used in this study were collected from 2014-2017.

Data was collected by placing a probe in the appropriate catheter holder and pressing it to the site of interest. A 3-dimensional volume is collected as the RPD spins and retracts the fiber through the protective sheath. Each pullback was collected in about 45s, depending on length of pullback (30mm-90mm) and pullback speed (1-10 mm/s). The rotational rate was set at 100 Hz; slower pullbacks result in higher resolution along the pullback axis. Based on clinical impression, the most pathologic site was selected and 1-10 pullbacks were performed; when possible, a contralateral pullback was also taken. If deemed clinically necessary, a biopsy was performed after image collection to confirm diagnosis, which could be co-registered to the OCT.

2.2.4. Common Artifacts in Endoscopic OCT

Air Bubbles

Imaging catheters are filled with water for refractive index matching between the silica optics and the plastic sheaths, as well as to allow for smooth motion of the catheter. However, this can introduce air bubbles into the optical pathway which obscure the tissue underneath (Figure 2.9). In the longitudinal view, bubbles are identifiable by vertical stripes of decreased intensity in the imaged tissue, coupled with high-intensity pixels within the protective sheath; bubbles may present in various sizes. As light travels faster through air than water, changes in the optical path length can be observed where the signal reaches the detector sooner and is captured as though the tissue has shifted towards the probe.

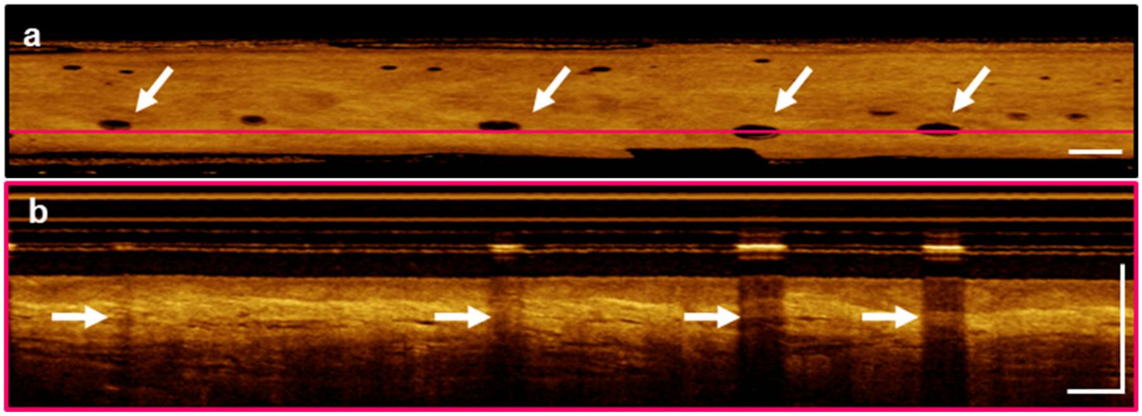


Figure 2.9. Effect of air bubbles in the imaging pathway, observed in the (a) en face view; (b) longitudinal view. Scale bars 1mm.

Sheath Markers

A second source of obstruction in the images are sheath markers, which are printed indicators on protective sheaths (Figure 2.10). These markers are uniformly distributed to allow for detailed localization of lesions of interest. Markers are identifiable by an approximate 0.5mm drop or loss of intensity in the tissue along the pullback axis. Unlike air bubbles, sheath marker obstructions are not accompanied by high intensity pixels.

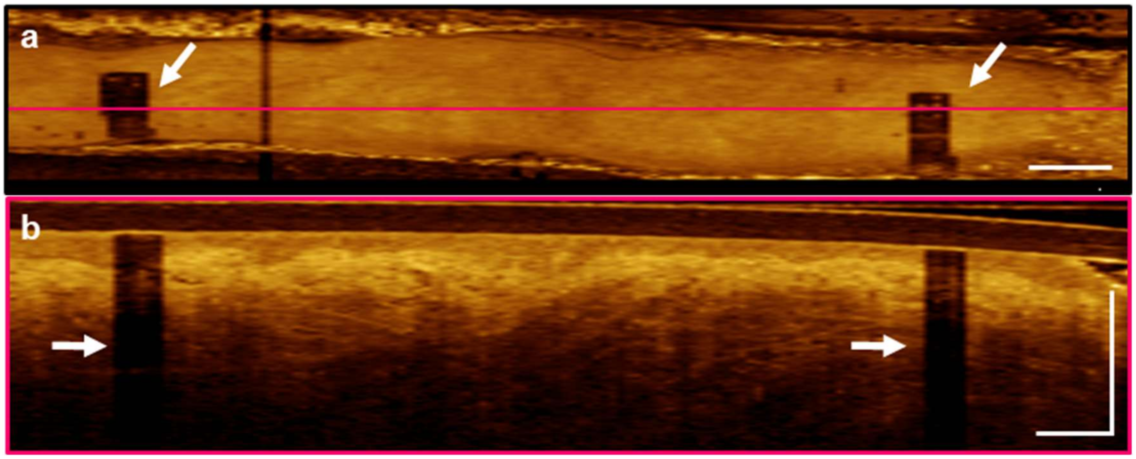


Figure 2.10. Effect of sheath markers in the imaging pathway, observed in the (a) en face view; (b) longitudinal view. Scale bars 1mm.

2.2.5. OCT of the Oral Cavity

OCT allows for visualization of tissue stratification across sites oral cavity, allowing for differentiation of the epithelial layer and the transition into the stromal layer, separated by the basement membrane. This is possible due to the changing backscattering properties caused by the unique cellular composition of each layer. In OCT, the epithelial layer is characterized by a darker layer, due to the lower scattering properties of this tissue. The stromal layer (demarcated by the basement membrane) is characterized by a brighter layer, due to the highly scattering properties of the tissue. These properties and labelled tissue structures are shown in Figure 2.11.

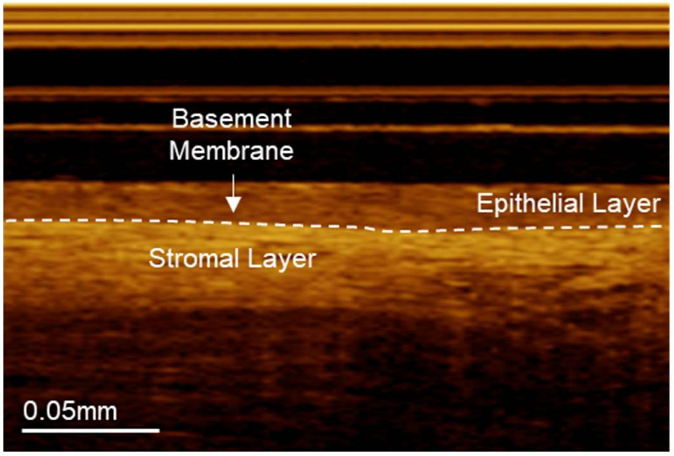


Figure 2.11. Normal layered morphology of the oral mucosa observed in OCT.

Discussed in section 2.1, epithelial layer thickness differs across tissue sites. Figure 2.12 details longitudinal slices of OCT pullbacks taken over healthy appearing samples of the buccal mucosa (panel a), ventral tongue (panel b), dorsal tongue (panel c), labial mucosa (panel d), lip (panel e), floor of mouth (panel f), vestibule (panel g) and gingiva (panel h). Imaging artifacts are noted by a white star. At all sites, OCT effectively captures changes in appearance and thickness of the epithelial layer. While changes in image intensity may be a result of differing scattering properties due to different epithelial thickness, tissue density or other properties, this information cannot be used reliably as changes in probe type, sheath catheter, adjustments to the reference mirror, among others, can also contribute to differences.

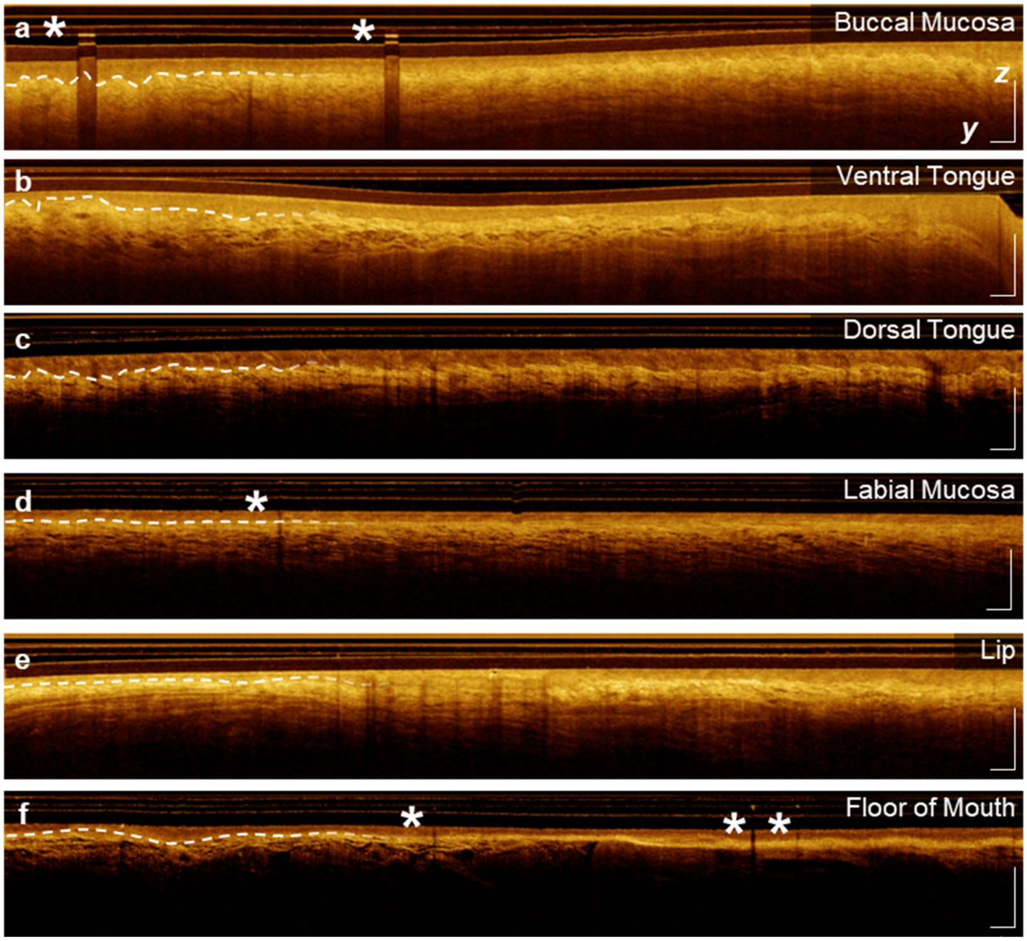


Figure 2.12. Longitudinal OCT scans of healthy appearing tissues of (a) buccal mucosa; (b) ventral tongue; (c) dorsal tongue; (d) labial mucosa; (e) lip; (f) floor of mouth; (g) vestibule; (h) gingiva. Annotations (white dash) have been included on the left of each slice to demarcate epithelial and stromal layers. White * denote imaging artifacts. Scale bars 1mm. Images have been stretched along the z-dimension to view details.

2.2.6. Cancer Detection using OCT

The BC Cancer Optical Cancer Imaging Lab has been previously demonstrated the utility of OCT as an optical biopsy device [6], providing insight into subsurface tissue structures which may indicate cancerous or pre-cancerous lesions. In OCT, healthy appearing tissue (Figure 2.13, panel a) is identified by clear and uniform stratification of the epithelial and stromal layers. Conversely, pathologic sites (OSCC example in Figure 2.13, panel b) are identified by thickening of the epithelial layer and subsequent destruction of the basement membrane, visible in OCT. Due to the nature of the data collected and used during this study, “healthy appearing tissue” or “clinically normal” is used to describe pullbacks that were taken across sites contralateral to the lesion of interest. Sites of interest were chosen through visual assessment by physicians in the clinic and biopsies were not performed on contralateral sites. Consequently, it cannot be confirmed that contralateral tissue imaged in this study is healthy; for example, many patients reported history of tobacco use, which is correlated with morphological changes across the entire oral mucosa [56]. Benign lesions supported by pathology are noted as such.

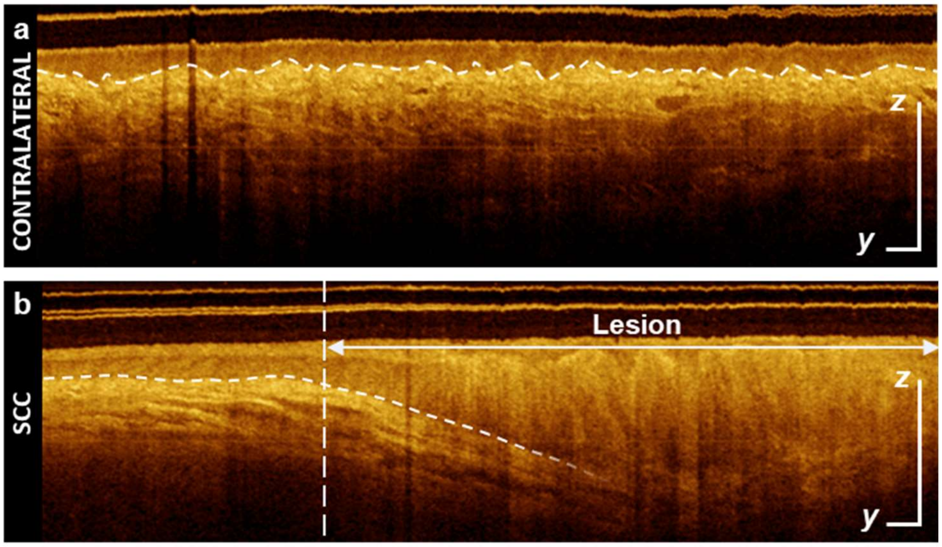


Figure 2.13. OCT of (a) clinically normal ventral tongue; (b) Biopsy confirmed OSCC diagnosis ventral tongue. Scale bars 1mm. Images have been stretched along the a-line (z) axis to view details.

2.3. Deep Learning

Neural networks (NNs) have influenced many modern technological advancements. However, the medical field has been hesitant to adopt artificial intelligence and NNs into clinical workflows. Poor uptake has been primarily attributed to lack of transparency and limited explanation of decisions [57].

Clinical adoption of OCT requires image analysis tools to provide rapid and reproducible assessment of tissue during biopsy procedures. However, the large volume of data collected with OCT makes manual annotation intractable. In this thesis, deep learning methods are explored as a tool to detect and quantify the presence and depth of the epithelial layer in OCT.

2.3.1. Neural Networks

Artificial NNs were developed with the goal of creating a computational approximation of how mammalian brains receive and processed information. In biology, neurons are excitable cells which carry electrical signals, integrating and processing information, and communicating with neighboring neurons and target cells [16]. In NNs, these features are mimicked with two computational elements: *nodes* and *connections*. Nodes receive and process information, while connections transmit data. Threshold stimulus required to initialize an action potential in biological neurons for communication is mimicked in NN by assigning '*weights'* to connections. Each node in a layer is connected to those in the subsequent layer, the sum of which denotes the importance of the node. A network where weights have been defined through prior training on a different dataset is referred to as *pre-trained*; continued training can be subsequently performed to *fine-tune* the network.

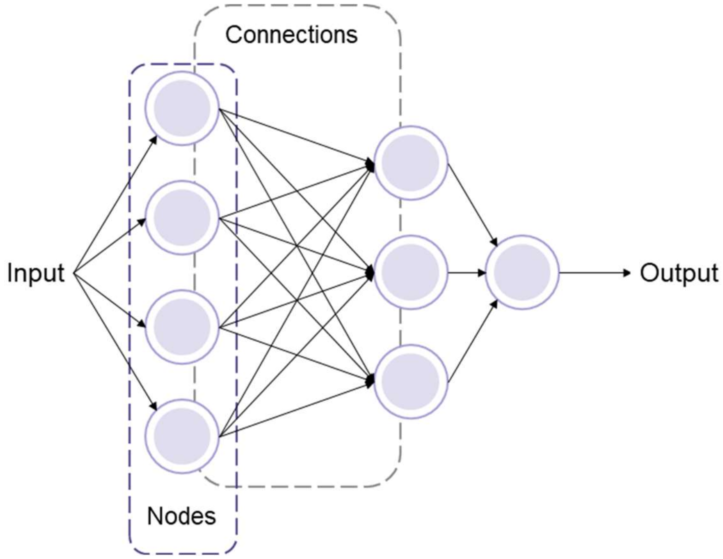


Figure 2.14. Connections and nodes of artificial NNs.

The universal approximation theorem suggests that for any function there exists a NN that converges on the exact or close approximation of the value [17]. Once trained, a given input is passed through a series of operations to formulate an appropriate output, commonly referred to as a network prediction.

2.3.2. Supervised Learning

Herein, supervised learning is implemented, where both data and reference information (ground truth labels) are made available to the network, and a loss function is used to assess the difference between the network predication and the reference information. This contrasts with semi-supervised or unsupervised learning, where the network is trained with limited or no reference information. NNs adjust their weights using an algorithm called backpropagation, which applies a criterion function in conjunction with an optimization function. The criterion (or loss) function quantifies the difference between the reference data and the predicted data, while the optimization function quantifies the weight adjustments that must be made to minimize the loss. A variety of loss and optimization functions are available; in image processing tasks, typical implementation uses the binary cross entropy (BCE) loss function (Equation 2.1), and the Adam [58] optimization function.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log\big(p(y_i)\big) + (1 - y_i) \cdot \log\big(1 - p(y_i)\big) \qquad 2.1$$

2.3.3. Terminology

The methods for networks applied in this research will be detailed further in Chapter 4. The following list briefly defines key terms that are used when defining NN training and implementation parameters.

- A *convolutional NN (CNN)* is a class of NN that specializes in image processing tasks, implemented as a series of convolutional layers, occasionally followed by a series of linear layers.
  - A *u-net* is a class of CNN that specializes in image segmentation tasks. It is implemented using a series of convolutional layers, called the encoder, followed by a series of deconvolutional layers, called the decoder; the connection point of the encoder-decoder is called the bottleneck. The output of each layer in the encoder is concatenated with the input of corresponding decoder layer, creating a u-like structure when viewed topographically.
- A *task* refers to a problem that the NN is trying to solve.
  - *Image classification tasks* assign entire images to a single class. The reference information for this type of task is referred to as a *target* or *label*.
  - *Image segmentation tasks* perform pixel-wise classification, assigning distinct image pixels to a category. In this work, instance segmentation is performed, which assigns all to instances of the same object to a single category, opposed to semantic segmentation, which allows for identification of unique instances. The reference information for this type of task is referred to as a *mask*.
- *Hyperparameters* are user defined variables initialized before training.
  - *Epochs* define the number of network iterations performed during training.
  - *Batch size* defines the depth of input data stacks loaded and shown to the network during each epoch. It is often used for memory optimization. An epoch is complete when all batches have been processed.

    - *Learning rate (LR)* adjusts the magnitude of the adjustments made during optimization. It may be updated during training to reduce overshoot and allow for precise identification of minimum error values.
    - *Initialization method* describes the algorithm used to initialize network weights.
    - *Regularization techniques* are methods that assist the network in generalizing and preventing overfitting. Techniques include early stopping, batch normalization and dropout.
- *Cohorts* are used to define the separation of data into various tasks for network development.
    - The *training cohort* is used to tune network weights and minimize the loss function.
    - The *tuning cohort* is used to evaluate the success of the current weight organization after each epoch. Minimal loss in the tuning cohort is an indicator that the network has effectively learned the task.
    - The *testing cohort* is evaluated once training has concluded and allows for evaluation of network on previously unseen data.
    - The *discovery set* defines the data used to build and evaluate the model (i.e. the training, tuning and testing cohorts).
    - The *external validation cohort* is ideally made up of data collected independently of the training/tuning/testing cohorts, for example through a different system or location. This designer should be blinded to this cohort and the external validation cohort should only be evaluated once the network is finalized to assess robustness.
- A *hidden layer* is an umbrella term that refers to all NN weight-based operations that are not the input or output layers. Adjustments to the organization and connection of node layers allows different features to be extracted and managed during training. Hidden layers relevant to this thesis are listed below.
    - *Linear Layers* are the basic building block of a NN, where each node in the previous layer is connected to each node in the current layer. Linear layers allow for a network to maintain precise information about data and make

connections with a lot of information. Linear layers are limited in ability to synthesize complex data and are computationally heavy.

- *Convolutional layers*, are linear layers which extract contextual and spatial information present in images. High level convolutional layers (earlier in the network topology) primarily identify coarse features with fewer filters, while lower level layers identify detailed features using many different filters.
- *Activation layers* are the only non-linear hidden layer. These layers are the backbone allowing NN to be universal approximators. In general, activation layers prevent nodes that do not meet sufficient thresholds from contributing to subsequent layers. Activation layers include the Rectified Linear Unit (ReLU) layer, and the sigmoid layer, where the thresholds are based on eq. 2.2 and eq. 2.3, respectively.

$$y(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{2.2}$$

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.3}$$

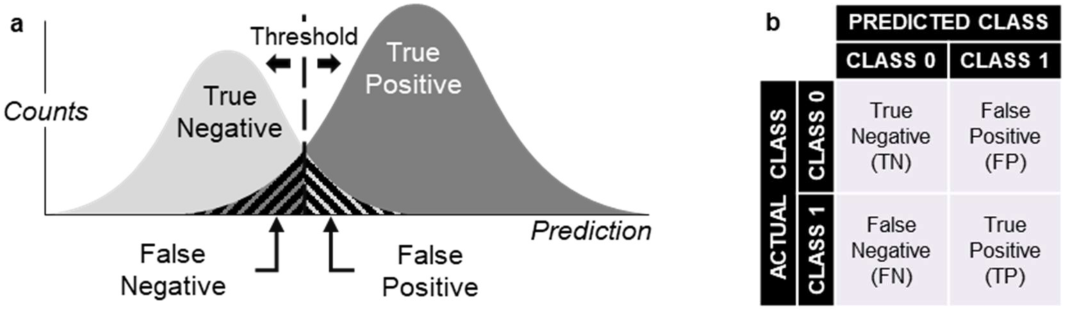
- *Pooling layers* are used to reduce image dimensionality while maintaining image integrity. Pooling information also benefits in reducing computational loads by reducing the number of weights within a network.
- *Concatenation layers* are merging layers that allow the network to combine weights into a single value, allowing for richer information content to be passed to subsequent layers.

2.4. Evaluation Metrics

For classification tasks, balanced accuracy, sensitivity, and specificity are reported, as well as the area under the ROC curve (AUC) and mean average precision (mAP) curves. For segmentation tasks, the dice similarity score (DSC) [59] and a custom metric adapted from mean average distance are reported.

2.4.1. Classification Metrics

In a perfect classifier the distribution of predictions is separated into two distinct groups: true positive (TP) and true negative (TN). In practice, these groups overlap and a threshold is applied to divide the results, creating false positive (FP) and false negative (FN) predictions. Careful selection of this threshold is necessary, as adjustments can bias a classifier. Selecting a threshold too far to the left will decrease the FN, but increases FP. Similarly, selecting a threshold too far to the right will optimize for FP at the cost of increasing FN error. FP misclassifications are referred to as type 1 errors, while FN misclassifications are referred to as type 2 errors. Figure 2.15 presents a visualization of this distribution (panel a) and a confusion matrix (panel b), commonly used to display counts of correct and incorrect predictions, after application of the classifier threshold.



		PREDICTED CLASS	
		CLASS 0	CLASS 1
ACTUAL CLASS	CLASS 0	True Negative (TN)	False Positive (FP)
	CLASS 1	False Negative (FN)	True Positive (TP)

Figure 2.15. Methods of visualizing classification distributions. (a) Bell curve distributions of classifier ability and effect of threshold selection; (b) confusion matrix allowing analysis of threshold effect.

Youden’s index (J) [60] is commonly applied in tasks aiming to equally minimize the count of FP and the FN predictions. Shown in eq. 2.1, this index is calculated from evaluating a range of threshold values and selecting the maximum value.

$$\mathrm{J} = \max\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} + \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} - 1\right) \qquad 2.1$$

From the confusion matrix, several metrics can be calculated to represent the predictive power of the classifier. In this thesis, sensitivity (also called recall; eq. 2.2), specificity (eq. 2.3) and balanced accuracy (eq. 2.4) are used. While the NNs defined in this work have binary targets, the nature of the tasks present imbalanced datasets. To better capture any bias, sensitivity and specificity are presented, which reflect the TP and TN rates (TPR, TNR), respectively. Balanced accuracy is also presented, which is an average of the specificity and sensitivity rates.

$$\text{Sensitivity} = \text{TPR} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad 2.2$$

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \qquad 2.3$$

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \qquad 2.4$$

Other metrics used to represent classifier success include the receiver operating characteristic (ROC) curve, and the precision-recall (PRC) curve. These metrics are threshold free, allowing for model-wide evaluation.

The ROC curve allows for visualization of the trade-off between the TPR (sensitivity), and the FP rate (1- specificity), as thresholds vary (Figure 2.16, panel a). A ‘better’ classifier will have a deeper curve, increasing in concavity with increased success. In contrast, a poor classifier will have a flatter curve, limited at the random guess model (linear y=x relation, dashed line in Figure 2.16, panel a). The area under the ROC curve (AUC, bounded between 0 and 1) is calculated from this plot, where a higher value represents a more successful classifier.

Conversely, the PRC plot (Figure 2.16, panel b) allows for visualization of the trade-off between precision (the positive predictive value, eq. 2.5) and recall (the sensitivity, eq. 2.2). Unlike the ROC curve, the PRC curve allows for a more complete analysis for models trained on datasets with imbalanced class representation [61], as is the case in this thesis. This invariance is possible as precision represents the proportion of accurate predictions (TP) within the positive predictions (TP + FP), while recall represents the ratio of positive predictions (TP) among all samples that should have been identified (TP + FN). The mean average precision (mAP) can be calculated from the area under the curve. A baseline classifier (dashed line, Figure 2.16, panel b) is often presented with in the plot, to allow comparison against a classifier that assigned all samples to a single class (i.e. a unskilled classifier). Accordingly, the value of the baseline classifier is dependant on the class distribution within the dataset. The equation for the PCR baseline curve is shown in eq. 2.6, where P represents the number of class 1 examples, while N represents the total number of examples.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.5}$$

$$\text{PCR}_{\text{baseline}} = \frac{P}{N} \tag{2.6}$$

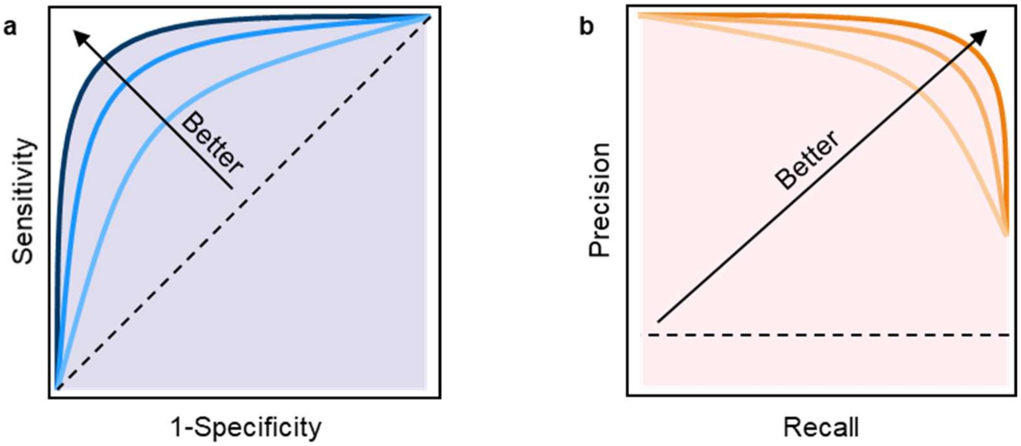


Figure 2.16. (a) ROC curves for improving classifiers (light → dark blue), AUC of best classifier (light purple region) and random guess model (black, dashed); (b) PRC plots for improving classifiers (light → dark orange), mAP of best classifier (light pink region), and baseline classifier (black, dashed).

2.4.2. Segmentation Metrics

Segmentation metrics aim to capture differences in pixel agreement by comparing reference masks to prediction masks. In semantic segmentation with a single object, the task is similar to binary classification, except prediction are per-pixel rather than the entire input. Now, a TP prediction defines a pixel that has been correctly identified as belonging to the mask, while a TN prediction defines a pixel that has been correctly identified as belonging to the background; FNs and FPs are pixels that have been incorrectly identified as belonging and not belonging to the mask, respectfully. In this work, the DSC (eq. 2.7, a common metric for DL segmentation tasks) is reported.

$$\text{DSC} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{2.7}$$

The segmentation tasks completed in this work are akin to a boundary detection task, which is currently a relatively novel field in DL, for which metrics have not been standardized. The DSC must be interpreted carefully in context of its bias towards larger segmentations. The error in a small shape (such as a thin boundary) is amplified in the DSC by the small overall sample size (pixel count). Conversely, large regions may exhibit high DSC despite containing numerous pixels that have been inaccurately segmented, as the sheer size of these regions overshadows the presence of errors. The bias towards larger shapes is depicted graphically in Figure 2.17, rows a and b. Despite Annotation 2 being offset by a single pixel from Annotation 1 in both cases, the larger objects result in a much larger DSC. At the extreme, a very small object can transition from a DSC of 100 to a DSC of 0 similar with a single pixel shift, depicted graphically in *rows c and d.*

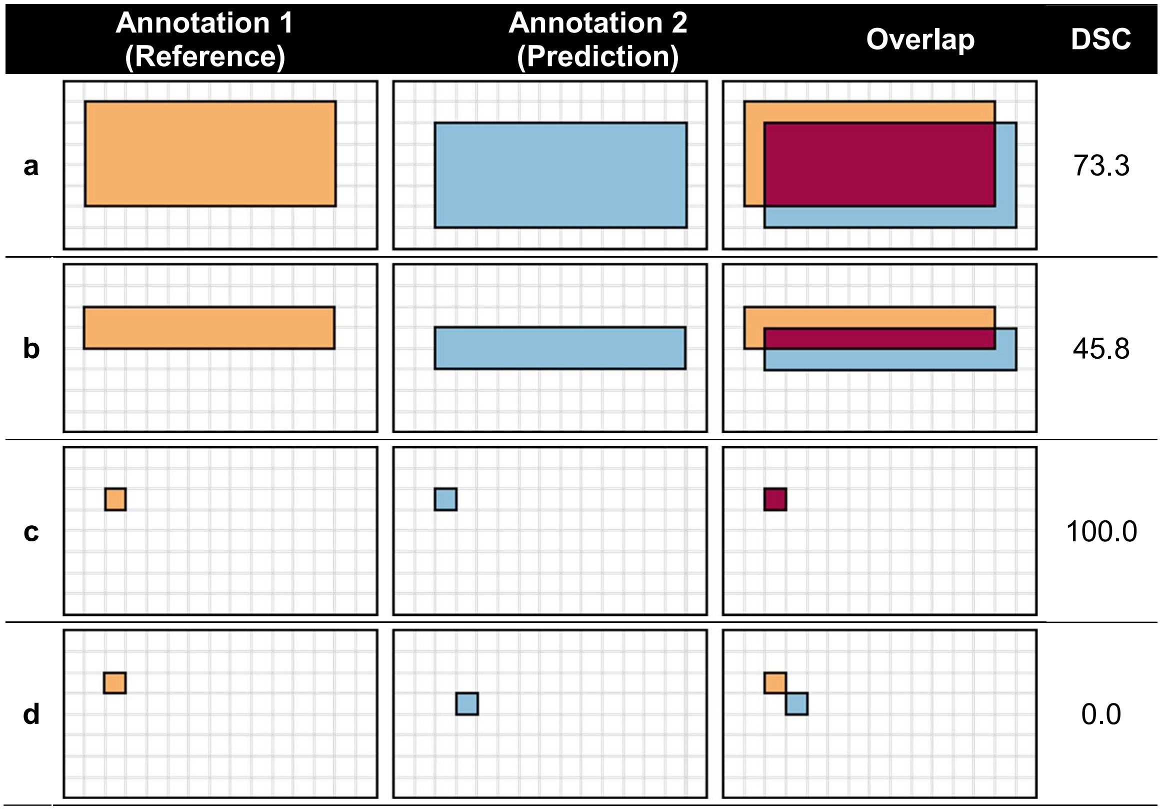


Figure 2.17. Pitfalls of the DSC, demonstrating that segmentation errors are more harshly evaluated in for smaller regions.
Adapted from [62]

In conjunction with the DSC, binary comparison metrics have been quantified to suit the primary aims of the segmentation tasks. While the DSC is useful in comparing similarity of objects, it does not account for regions correctly identified to contain no objects, an important detail for this work. Moreover, as the segmentation tasks discussed in this thesis are closer to boundary detection problems, quantifying the distance between predictions along the a-line direction is also important. Thus,

1. The proportions of specific agreement (positive agreement [PA] and negative agreement [NA]) are reported to quantify regions of agreement regarding the existence/lack of a boundary [63].

2. Mean (μ) and standard deviation (SD, σ) metrics are calculated to quantify the distance between predictions in regions where both annotations have identified a boundary.

Similar to classification tasks, a confusion matrix can be built to quantify accuracy of annotation presence. In this application, each a-line is classified as containing a boundary or not. A TP occurs when both annotations contain a boundary, while a TN occurs when neither contain a boundary. FPs and FNs occur when only one rater classifies the a-line as containing a boundary. From the confusion matrix, many metrics can be calculated, including the commonly presented proportion of observed agreement (Po, eq. 2.8).

$$Po = \frac{TP + TN}{TP + FP + FN + TN} \qquad 2.8$$

However, this metric fails to properly represent model performance in cases with imbalanced data. Instead, PA (eq. 2.9) and NA (eq. 2.10) are reported to measure percent agreement between the presence of the boundary, and percent agreement of no boundary; the PA is sometimes referred to as the F1 Score. Reporting agreement metrics such as Po, PA, and NA is preferable over relative metrics (such as Cohen’s Kappa) as it does not require the reader to interpret the metric.

$$PA = F1\,Score = \frac{2 * TP}{2 * TP + FP + FN} \qquad 2.9$$

$$NA = \frac{2 * TN}{2 * TN + FP + FN} \qquad 2.10$$

While the PA and NA metrics compares rater agreement regarding to whether a boundary exists, they do not inform on how far away the areas of agreement may be. To account for positional disagreement, mean and SD metrics are evaluated on a vector generated by the differences in mask location along the pullback axis for each longitudinal slice. Shown in eq. 2.11, *r* represents the row where the boundary has been defined for annotation *a*, along the c^{th} column.

$$\mathrm{Diff} = |\mathrm{r}_{\mathrm{a}_1,\mathrm{c}} - \mathrm{r}_{\mathrm{a}_2,\mathrm{c}}| \qquad 2.11$$

Figure 2.18 presents a graphical summary of the conditions used to generate the confusion matrix, with the inset demonstrating measurement of pixel distance between the prediction and reference at each a-line along the pullback axis.

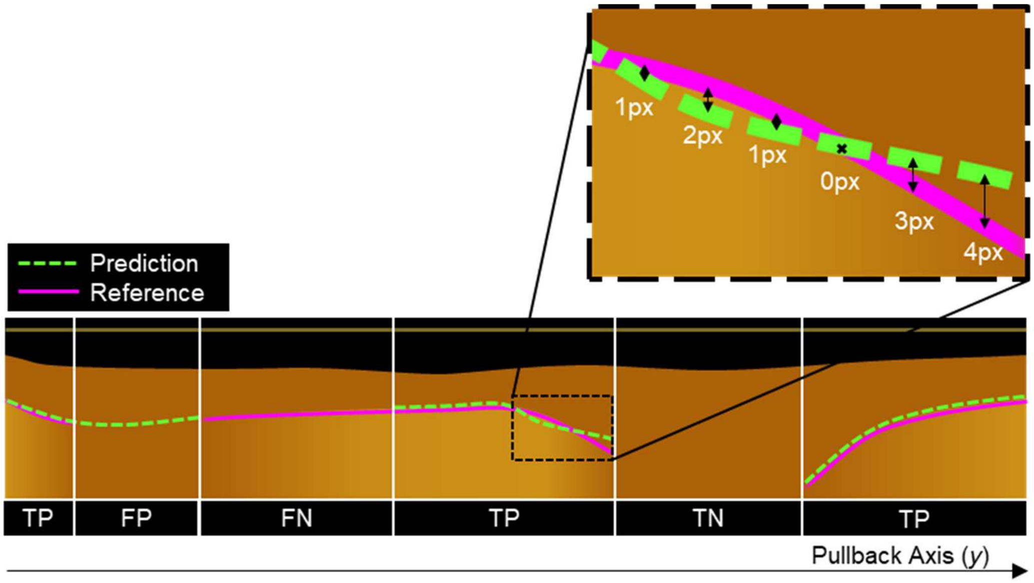


Figure 2.18. Graphical example of areas of overlap for segmentation metrics, with inset demonstrating pixel distance metric.

Chapter 3. Reference Data Generation

To our knowledge, we are the first group to develop automated segmentation algorithms of oral OCT data using DL tools; the novelty of this work necessitated the development of a reference data set to train and evaluate DL networks. Methods for dataset selection, rater training and generation of reference data are described below, as well as a summary of the MATLAB app developed to facilitate annotation collection.

3.1. DL Dataset Selection

From the data collected during the study described in Section 2.2.3, 184 pullbacks were selected, from 66 patients (37 male, 28 female, 1 missing data). Selections were made through qualitative and quantitative assessment. Visual analysis was used to ensure good OCT quality and minimize presence of imaging artifacts (could not make up more than approximately 40% of entire slice). Pullbacks were also required to be >30mm, with pullback speeds ≤10mm/s). Images were selected to encourage network robustness and generalizability by including a variety of imaging sites and pathologies. Up to 3 longitudinal slices were selected from each pullback, with at least 15° of separation between slices. Table 3.1 and Table 3.2 summarize the distribution of pathologies and imaging sites, respectively. An unknown lesion label is used to describe pullbacks that were taken across a clinically suspicious lesion, but subsequent biopsy was not taken/available.

Table 3.1. Summary of pathology in DL dataset.

Diagnosis	No. Pullbacks	% of Dataset	No. Longitudinal Slices	% of Dataset
Contralateral	**57**	**31.7**	**126**	**44.5**
Other	**14**	**7.6**	**21**	**7.3**
Benign (pathology confirmed)	5		8	
Actinic Cheilitis	2		3	
Candidiasis	4		6	
Scar	3		4	
Hyperplasia/Dysplasia	**47**	**25.5**	**60**	**20.8**

Hyperplasia	2		2	
Verrucous Hyperplasia	3		5	
Dysplasia Grade 1	20		26	
Dysplasia Grade 2	15		17	
Dysplasia Grade 3	7		10	
In Situ Cancer	**5**	**2.8**	**7**	**2.5**
CIS	5		7	
Metastatic Cancer	**25**	**13.9**	**29**	**10.2**
VC	8		8	
OSCC	17		21	
Unknown Lesion	**36**	**20**	**45**	**15.9**
Total	**180**	**100.0**	**283**	**100.0**

Table 3.2. Summary of dataset site selection

Site	No. Pullbacks	% of Dataset	No. Longitudinal Slices	% of Dataset
Buccal Mucosa	22	12.0	34	11.8
Floor Of Mouth	8	4.3	13	4.5
Gingiva	10	5.4	11	3.8
Labial Mucosa	5	2.7	11	3.8
Lip	1	0.5	1	0.3
Tongue - Dorsal	4	2.2	6	2.1
Tongue - Lateral	56	30.4	86	29.9
Tongue - Ventral	72	39.1	119	41.3
Vestibule	6	3.3	7	2.4
Total	**184**	**100.0**	**288**	**100.0**

3.2. Data Pre-Processing

The following pre-processing steps were performed on each longitudinal slice to ensure uniformity, with results shown in Figure 3.1 below.

1. Resampling such that each pixel was 10µm (panel a).

2. Filtering with a 2nd order boxcar filter (averaging preceding and proceeding two slices, total of five slices, panel b). High data collection speeds meant that neighboring longitudinal slices contain near-identical features, but differences in image noise would be reduced through averaging.

3. Remapping intensity values using MATLAB's `imadjust` [64] function (panel c), such that the minimum intensity was the mean of the noise floor (representing the background noise of the system; selected from the deepest row of the slice). The maximum remained unchanged.

Pre-processed images were saved as .tif files with 16 bit-depth resolution.

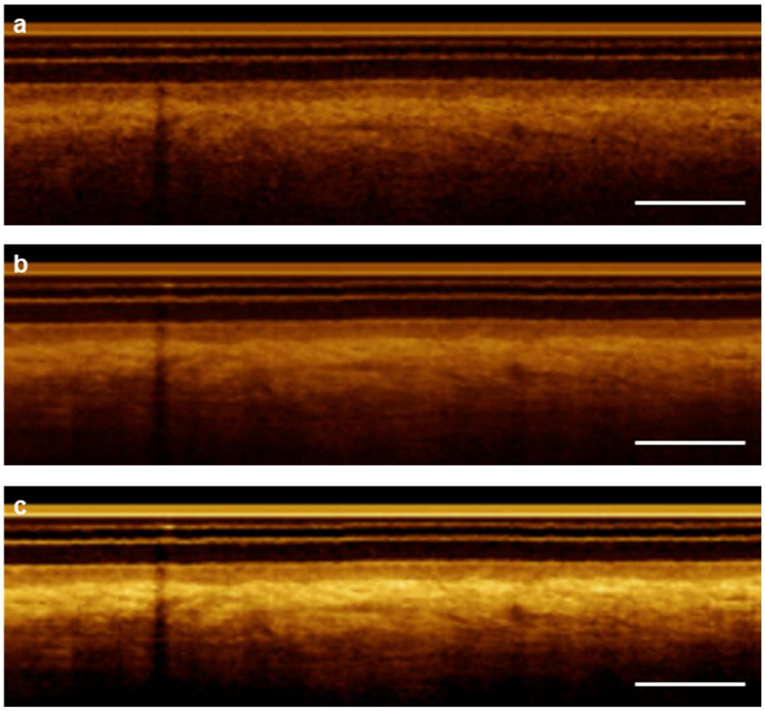


Figure 3.1. Longitudinal slice pre-processing steps; (a) resampled image; (b) filtered image; (c) balanced image. Scale bars 1mm.

3.3. MATLAB Software Development

Current annotation building software was deemed unsuitable for creating references of *in vivo* oral OCT, as primary operation required users to trace lines across the entire tissue surface using a computer mouse. In lieu of this, I developed an in-house annotation software, called SegApp.

3.3.1. Functional Requirements

SegApp allows users to view and generate image annotations. Requirements and justifications for the development of SegApp are summarized in Table 3.3.

Table 3.3. Functional requirements for development of SegApp Matlab Application

	Requirement	Justification
1	Load .jpg, .tif and .png image files	Utility for common image file types
2	Compatible with MATLAB R2019a and later	-
3	Line segments are defined by mouse-click *control points*	Defining line segments through control points allows for precise localization of key points and reduces off-target errors typical of mouse-drag line drawings
4	Interpolated segments defined by control points are displayed in real time	Allows for accurate visualization of line segments
5	Control points are adjustable through mouse drag; control points are deletable; control points can be added to existing line segments	To easily accommodate off-target mouse clicks
6	Line segments are labelled; custom labels can be generated	Allows for task generalization and detection of various regions of interest (ROIs)
7	Line segments may contain up to 100 control points; images may contain up to 50 line segments	Intricate features require detailed line segments
8	Line segments can be saved, reviewed, adjusted and deleted	Allows user to return to previously generated line segments, review and edit as needed
9	Navigation must be possible while generating line segments	Large images require zoom to view features and out-of-slice information should be accessible without changing operational mode

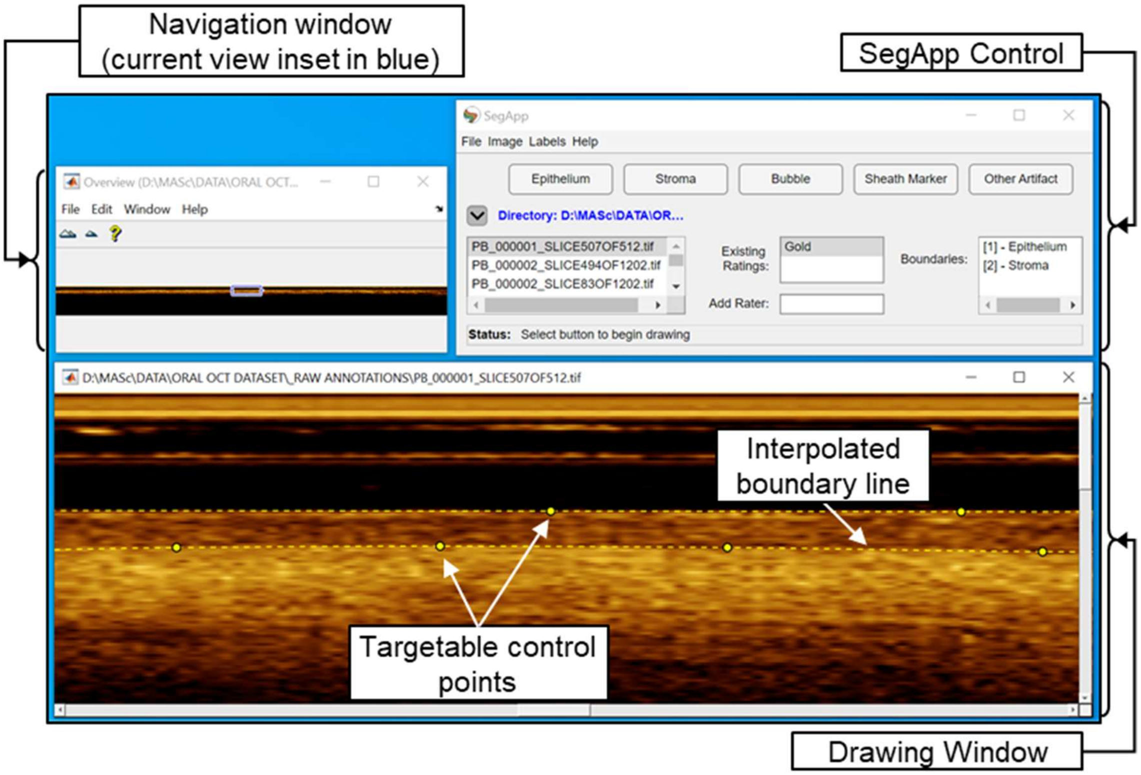
3.3.2. Implementation and Use

SegApp is implemented with (1) an independent control GUI window, (2) a targetable frame to display the image (‘drawing window’), and (3) a window of the entire image (‘navigation window) (Figure 3.2). The drawing window can be navigated (zooming, panning) using keyboard or mouse input. The navigation window displays the current position and the drawing window inset may be dragged for quick adjustment of the current drawing window view. The inclusion of these tools was necessary to accommodate the high aspect ratio OCT data, which could be up to 9000 pixels wide, and 382 pixels deep.

To view existing annotations, the user must select a directory containing the image and annotation stacks that they wish to review. The local image stack is loaded into a targetable list box and any annotations associated with the current image populate a second list box ('Existing Ratings'), designated by rater ID. If a rater ID is then selected, all annotations tagged by the selected rater within the selected image slice are displayed in the 'Boundaries' list box. These boundaries are selectable, which triggers the highlight of the corresponding annotation in the drawing window.

To create a new annotation, the user must enter their initials into 'Add Rater' text box. Users can then add annotations to the image by selecting a label button and using the computer mouse to generate 'control points' by clicking on regions of interest. A boundary line is interpolated and displayed in real time using MATLAB's `pchip` [65] algorithm. Each control point can be edited or deleted, and each boundary line can be re-labelled, deleted or fine-tuned with more control points.

A summary of available controls and features is detailed Figure 3.3. Additional features include customizable colormaps for greyscale images, addition or removal of labels, a built-in user manual and a keyboard shortcuts dictionary.



Directory: D:\MASc\DATA\OR...

PB_000001_SLICE507OF512.tif
PB_000002_SLICE494OF1202.tif
PB_000002_SLICE83OF1202.tif

Existing Ratings: Gold

Boundaries: [1] - Epithelium
[2] - Stroma

Status: Select button to begin drawing

D:\MASc\DATA\ORAL OCT DATASET_RAW ANNOTATIONS\PB_000001_SLICE507OF512.tif

Figure 3.2. Overview of SegApp user interface. SegApp Control expanded in Figure 3.3.

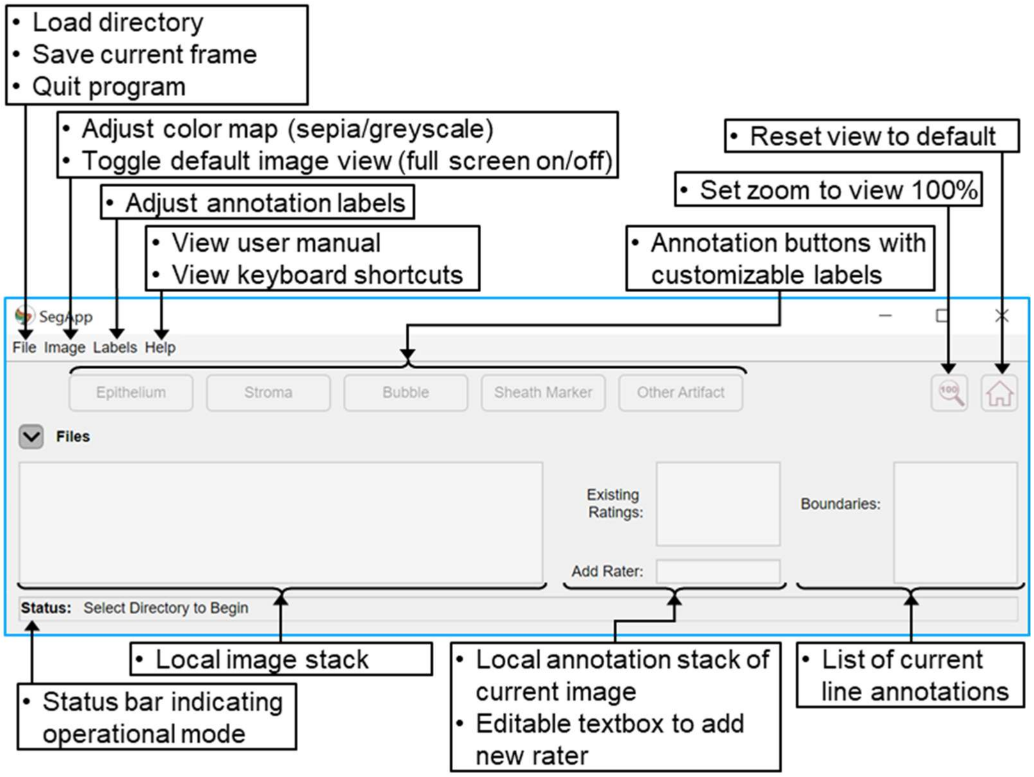

• Load directory
• Save current frame
• Quit program
• Adjust color map (sepia/greyscale)
• Toggle default image view (full screen on/off)
• Adjust annotation labels
• View user manual
• View keyboard shortcuts
• Annotation buttons with customizable labels
• Set zoom to view 100%
• Reset view to default
SegApp
File Image Labels Help
Epithelium
Stroma
Bubble
Sheath Marker
Other Artifact
Files
Existing Ratings:
Add Rater:
Boundaries:
Status: Select Directory to Begin
• Local image stack
• Status bar indicating operational mode
• Local annotation stack of current image
• Editable textbox to add new rater
• List of current line annotations

Figure 3.3. Detailed overview of SegApp control interface.

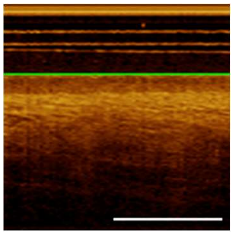
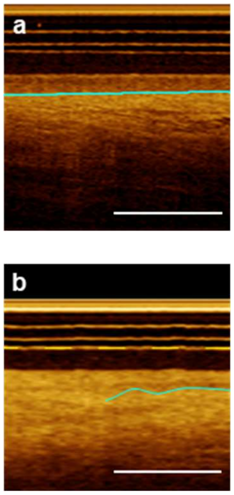
3.4. Manual Annotation

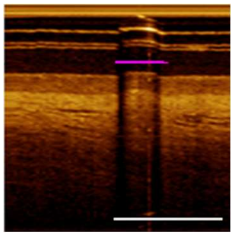
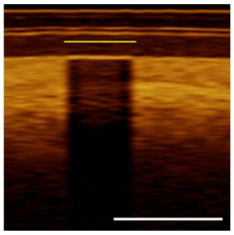
For this research, manual segmentation of the (1) epithelial surface, (2) basement membrane, (3) bubbles and (4) sheath markers was completed using SegApp. Annotations were generated by 6 raters (five experienced OCT researchers, one undergraduate trainee). Raters were initially distributed a training package and subsequently assigned a portion of the selected dataset. Data was de-identified, and raters were blinded to all patient data, including diagnosis and imaging site.

3.4.1. Rater Training

An in-person training session was provided and a training package was compiled and distributed, containing descriptions and examples of landmarks to identify the four regions of interest, summarized in Table 3.4. Included in Appendix A, the training package of 10 images, consisting of eight ‘easy’ slices and two ‘difficult’ slices. Slices were considered ‘easy’ if they had clear, uniform tissue stratification, and ‘difficult’ if they demonstrated a loss of resolvable stroma or substantial artifacting. A standardization session reviewing rater impressions was completed all raters had concluded their training packages.

Table 3.4. Oral OCT landmark descriptions used to create reference annotations

Landmark	Description	Sample Annotation (scale bars 1mm)
Epithelial Surface	The exterior layer of the tissue. *Confounding Factors*: Uneven and inconsistent layers of window tube and protective sheaths, air bubbles and sheath markers, mucous, or keratinization at the tissue surface.	
Basement Membrane	The transition of the epithelial layer to the stromal layer. In healthy appearing tissue, it can be identified by a sharp intensity transition from the dark epithelial layer to the bright stromal layer. In dysplastic tissue, this transition becomes more gradual, but the basement membrane was identified as the region of highest change in intensity gradient (Fig. a). *Confounding Factors*: As tissue progresses into more pathologic conditions, the transition may disappear completely. Raters were instructed to leave this area blank (Fig. b). Other confounders include bubbles and sheath markers, presence of sub-epithelial ducts, tissue folds and air gaps.	

Bubble Artifact	Air bubbles in imaging path. Identified as vertical stripes of decreased intensity in the imaged tissue, coupled with high intensity pixels within the sheath. Bubbles may present in various sizes. Raters were instructed to generate a horizontal segmentation that covered the length of the artifact. *Confounding Factors*: May appear similar to sheath markers (below).	
Sheath Marker Artifact	Printed indicators on sheaths. Identified by an approximate 1mm decrease or loss of intensity in the tissue, but not accompanied by the high intensity pixels characteristic of bubbles. Raters were instructed to generate a horizontal segmentation that covered the length of the artifact. *Confounding Factors*: May appear similar to bubbles (above).	

3.4.2. Reference & Consensus Set Generation

To ensure image features were consistently annotated and reduce rater bias, three raters were randomly assigned to generate annotations for each of the 288 longitudinal slices. Cases that raters identified as difficult, or cases with significant inter-rater disagreement were discussed in a panel review setting with all raters present to select the best annotation (examples shown in Figure 3.4).

During review sessions, some key issues were identified; a primary issue was experience in parsing OCT data, where further training would have been useful in creating accurate segmentations (Figure 3.4, panel a – yellow rater selected as the most correct). More nuanced issues included the level of zoom that each rater used to generate their annotations, where insufficient zoom would result in drift due to mouse click resolution, while too much zoom resulted in loss of spatial context. There was also expected differences in interpretation of structures (Figure 3.4, panel b), where it became very difficult to decide which annotation was correct.

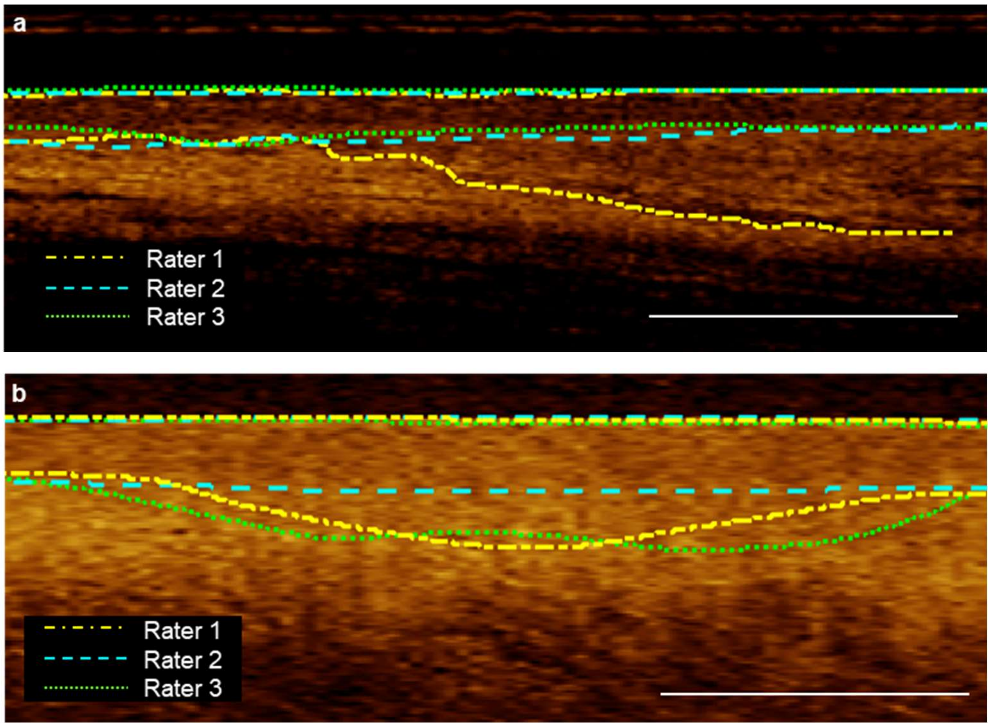


Figure 3.4. Examples of inter-rater disagreement; (a) disagreement due to lack of experience, (b) disagreement due to nebulous stratification. Scale bars 1mm

Fully supervised NNs are limited to a single reference for each input image, thus rater impressions were combined into a single reference ground truth. This was a challenging problem as the annotations were lines with single pixel thickness. Multiple methods were explored to combine annotations, including pixel voting, rater averaging and STAPLE [66]. Each of these methods presented unique drawbacks, but all resulted in incorrect annotations or loss of data and required extensive manual corrections. In-lieu of combining annotations from each rater, a consensus dataset was created a single rater using existing impressions as a reference. To avoid bias in difficult cases, an additional review meeting was held to allow all 6 raters to provide feedback. To reduce mouse-click drift and ensure sufficient click-point resolution, the zoom was set at approximately 200%. An example of a complex case is shown in Figure 3.5, panel a. Rater impressions demonstrating inter-rater disagreements are overlaid in panel b, and the consensus annotation is overlaid in panel c. The unknown sub-epithelial structure (speculated to be a duct) may be compressing cells in the area and creating a region of increased scattering, mimicking the transition at the basement membrane.

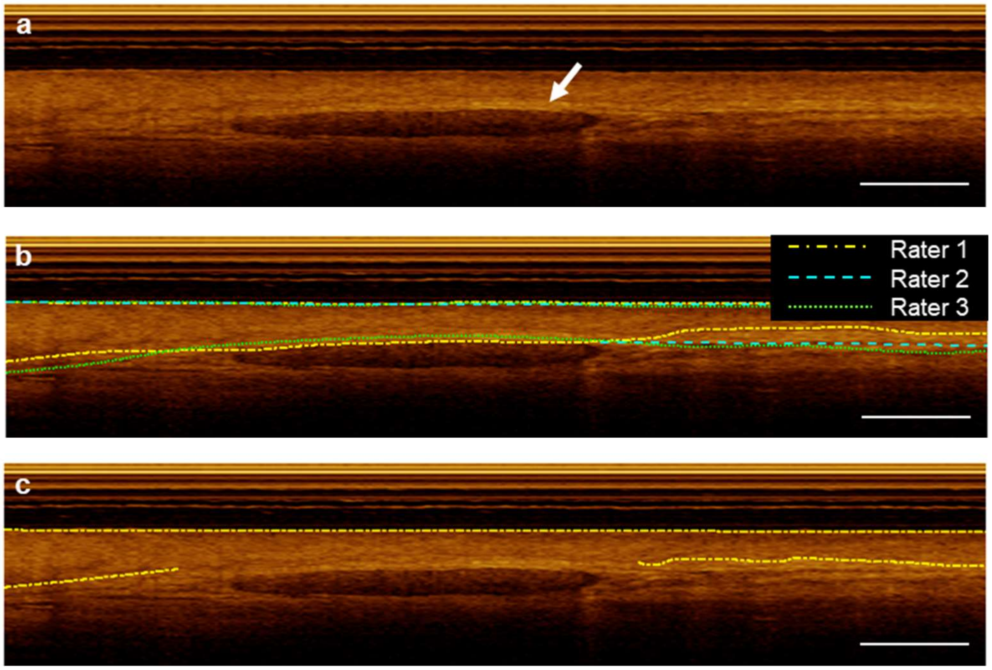


Figure 3.5. (a) OCT image of a complex case, confounded by the presence of a sub epithelial structure (white arrow); overlaid with (b) inter-rater disagreement; (c) consensus reference used for DL training.

Cases such as this lend an ambiguity to the precise location of the tissue transitionary zone. Consequentially, the reference set generated for this thesis can be referred to as a ‘noisy’ reference, meaning it may contain errors or present with inconsistent information. The vastness of the dataset, compounded with inter and intra rater disagreements provide evidence that it is difficult to manually generate accurate segmentations boundaries and further motivate the need for an automated system.

Chapter 4. Epithelial Layer Segmentation with Deep Learning

This chapter summarizes the cohort selection, network development, and training and testing methods of four independent NNs designed to segment the epithelial layer in an oral OCT volume. The final output this pipeline is an enface map encoding epithelial thickness.

Analogous to the boundaries that raters were asked to generate during annotation creation, three discrete tasks were identified for automated processing: segmentation of the epithelial surface, segmentation of the basement membrane, and identification of imaging artifacts. The thickness of the epithelial layer can be represented as a depth map, which effectively encapsulated information about the presence of the basement membrane and the thickness of the epithelial layer. This depth map could subsequently be superimposed onto the *en face* projection of the pullback, facilitating localization of areas of increased epithelial thickness and of image artifacts.The helical scanning pattern of the OCT system presented an additional processing task of detecting the FOV containing good probe to tissue contact.

Summarized in Figure 4.1, an automated pipeline was developed to accomplish the prior defined objectives. Pre-processing methods included contrast and brightness balance (described in Section 3.2), as well as image partitioning. Further detailed in Section 4.2, the first DL network is trained to discard images that were not within the imaging FOV. Subsequent networks were implemented to segment the tissue surface (described in Section 4.3), basement membrane segmentation (described in Section 4.4) and identification of confounding artifacts (described in Section 4.5). These networks could be run in parallel, but were implemented sequentially due to computational constraints. Post-processing steps are unique to individual networks and are described within each section when necessary, and cumulate in epithelial layer maps providing information about changes to epithelial thickening and stratification (described in Section 4.6). Pre- and post-processing methods are completed in MATLAB, DL networks are implemented with PyTorch framework using NVIDIA Cuda v11.4, and coded in Python 3.6.9. All experiments are performed on a Windows 10 operating system, with CPU Intel Core i7-4770 3.40 GHz, GPU NVIDIA GeForce GTX 1660, and 32 GB of RAM.

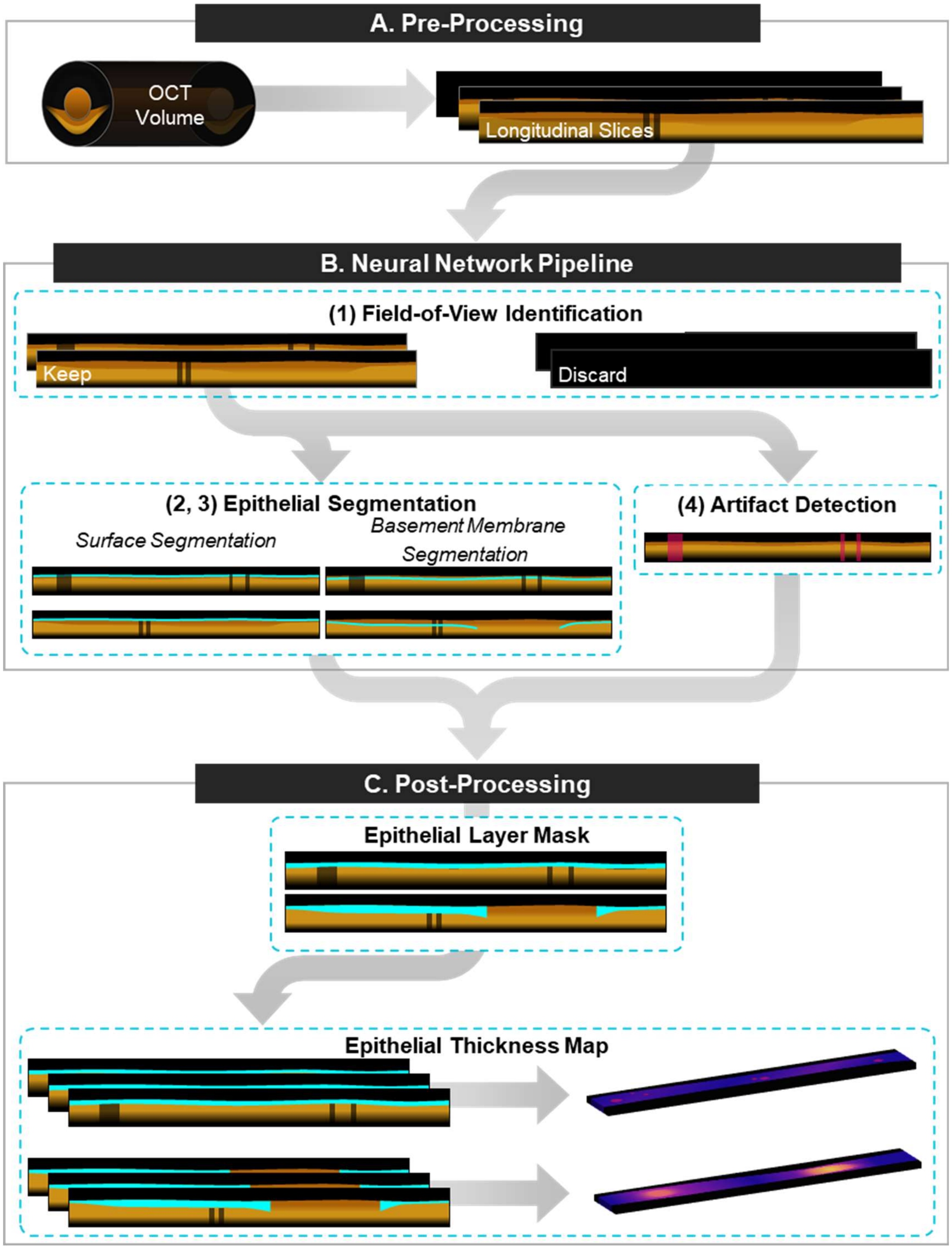


Figure 4.1. Automated OCT processing pipeline. Demonstrates flow from pre-processing steps to DL networks (includes (1) detection of the imaging field, (2) segmentation of the epithelial surface, (3) segmentation of the basement membrane, (4) detection of imaging artifacts), and completes with construction of epithelial thickness map. Schematics of contralateral and pathologic longitudinal slices are depicted when informative.

Subsequent sections will discuss the methods (dataset preparation, network selection and training protocols, and post-processing methods), results, and discussion (including representative examples) for each of the DL networks. Methodology common to all DL networks is presented in the next section.

4.1. Common Methodology

DL networks are trained from scratch for each task, but separate networks may share training hyperparameters. Specific hyperparameters for the FOV classification network, epithelial surface segmentation networks, basement membrane segmentation network and classification of imaging artifact presence network are defined in Table 4.1.

Batch size and number of epochs are established through experimentation; batch size is also limited by input image shape. The initialization method, criterion and optimizer are typical for the task. The PyTorch default weight initialization method was used, which creates a uniform distribution bounded by $\sqrt{\mathrm{No.\,Features}}^{-1}$.

An LR scheduler was implemented to decrease the LR by a factor of 0.1. The patience (count of epochs without loss improvement) was established through hyperparameter experimentation, as was the minimum LR.

Early stopping was implemented as a regularization technique, and triggered if the loss does not improve by the defined threshold within the number of epochs defined by the patience. When triggered, the weights from the epoch with the lowest tuning loss are used.

Image augmentation was applied to increase the amount of training data, with 50% likelihood of horizontal image flips occurring, and up to 10% image shift along the vertical axis. These parameters were identified as representative of variations occurring during OCT data collection, without introducing non-representative data.

A pixel classification threshold was defined for the segmentation tasks, which served to binarize the output of the final layer and create mask predictions.

Table 4.1 Training hyperparameters for DL network development, highlighting shared parameters for FOV (classification network), Epithelial Surface (segmentation network), Basement Membrane (segmentation network) and Imaging Artifacts (segmentation network).

Hyperparameters	FOV	Epithelial Surface	Basement Membrane	Imaging Artifacts
Number of Epochs	10	30	20	30
Batch Size	64	8	8	64
Initialization Method	uniform distribution			
Criterion	BCE [67]			
Optimizer	Adam			
LR	$1x10^{-4}$			

Scheduler: Reduce LR on Plateau	patience = 3,
	factor = 0.1,
	min. LR = $1x10^{-8}$
		patience = 5,
		factor = 0.1,
		min. LR = $1x10^{-8}$
			patience = 5,
			factor = 0.1,
			min. LR = $1x10^{-8}$
				patience = 3,
				factor = 0.1,
				min. LR = $1x10^{-7}$

Dropout	0.5	0	0	0.5

Early Stopping;	patience = 5,
	Δ_{min} = 0.01
	mode = min. loss
		patience = 20,
		Δ_{min} = $1x10^{-4}$,
		mode = min. loss
			patience = 10,
			Δ_{min} = 0.01,
			mode = min. loss
				patience = 5,
				Δ_{min} = 0.001
				mode = min. loss

Augmentation	horizontal flip 50%, y-axis shift ± 10%			
Pixel Classification Threshold	N/A	0.5	0.5	N/A

4.2. Field-of-View Detection

This section describes the development and implementation of a network developed to identify regions within the imaging field of view. The 3D imaging volume generated by the helical scanning pattern of the OCT system is well suited for luminal organs such as the small airways of the lungs or fallopian tubes, images of the oral cavity produce 25-40% of the longitudinal slices absent of oral tissue, and 10-20% of slices with tissue surface suffer from poor surface contact. As future steps in the segmentation workflow expect slices that contained good probe contact, the first step of the pipeline was to exclude all regions that existed outside of the tissue FOV.

4.2.1. Methods

Dataset Preparation

Training, tuning, and testing cohorts were generated by selecting longitudinal slices from 9 pullbacks. Only slices with complete tissue contact (class 1, n = 2204) or no tissue contact (class 0, n = 1545) were selected. To reduce manual annotation requirements, slices with partial or poor contact were excluded from training (n = 1694). Figure 4.2 presents sample longitudinal slices of class 1 (panel c), excluded (panel d) and class 0 (panel e), as well as their respective locations within a cross-section (panel 1) and *en face* view (panel b).

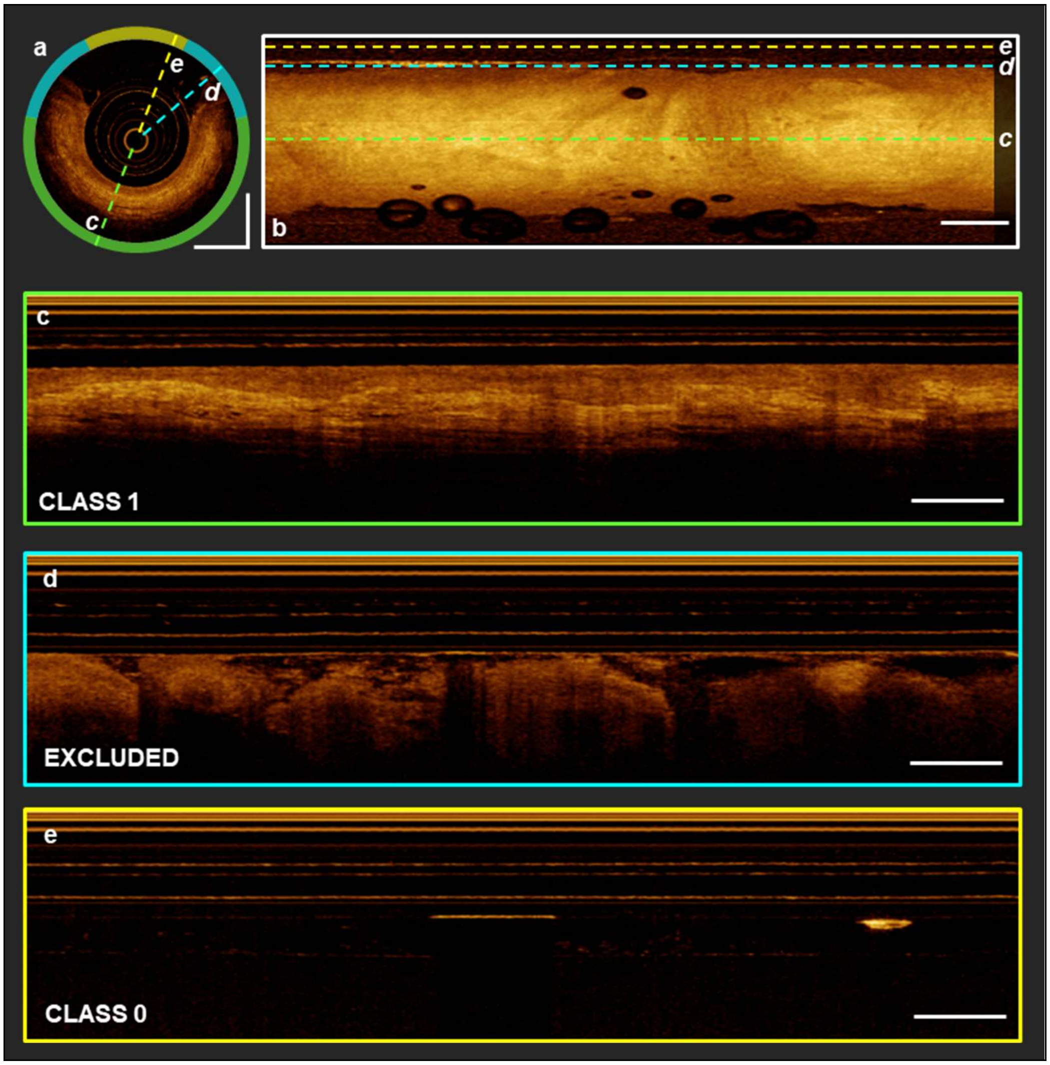


Figure 4.2 Helical scanning OCT with exclusionary criteria for FOV detection. (a) Cross-sectional and (b) en face view with class 0 (yellow), excluded (blue) and class 1 (green) slice markers; (c) (d) (e) represent FOV class 1, excluded and class 0, respectively. Scale bars 1mm.

Each longitudinal slice was divided into 256x256 pixel tiles, with 128 pixel overlap, generating 152,158 tiles; original slices were 4000-9000x382 pixels. Partitioning the image eliminates the need for downsampling and created memory-conscious, uniformly sized inputs as required by NNs. Additionally, image tiling produced more samples with which to train the network, opposed to a single longitudinal slice. The distribution of tiles for network development is summarized in Table 4.2. All cohorts were normalized using z-score normalization to reduce outlier influence [68].

Table 4.2. Distribution of tiles for FOV network development

Cohort	No. Tiles (% of cohort)		No. Patients
	Class 0	Class 1	
Train	49430 (42.3)	67464 (57.7)	7
Tune	7790 (43.8)	9994 (56.2)	1
Test	7030 (41.5)	10374 (58.5)	1
Total	64250	87376	9

Network Definition & Parameters

A custom CNN was used to train the FOV detection network. A shallow network with two convolutional layers was chosen to reduce overfitting. Network topology is shown in Figure 4.3. Further details on network layers can be found in Appendix B, Table B1. Hyperparameters implemented for network training are summarized in Table 4.1.

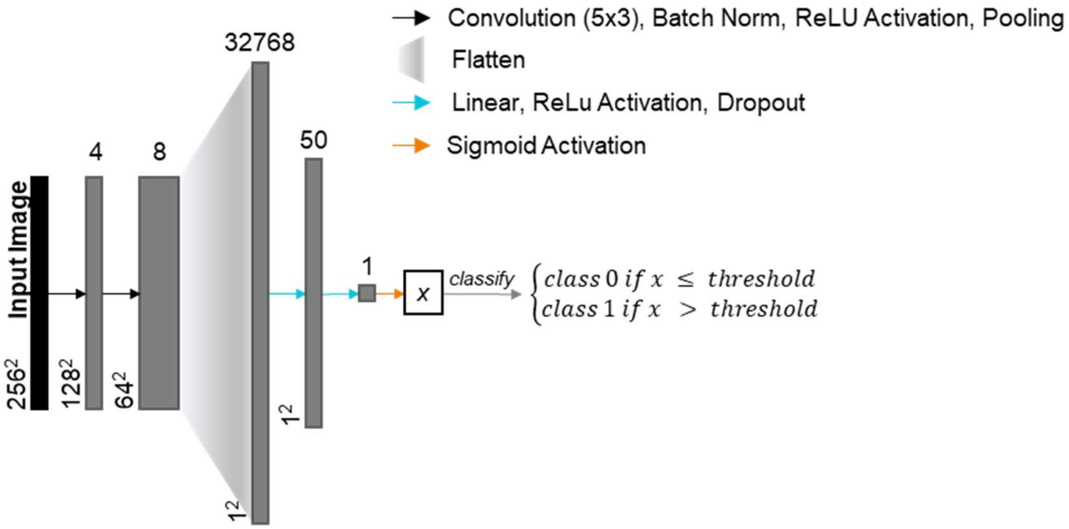

Convolution (5x3), Batch Norm, ReLU Activation, Pooling
Flatten
Linear, ReLu Activation, Dropout
Sigmoid Activation
Input Image
256^2
4
128^2
8
64^2
32768
1^2
50
1^2
1
x
classify
$\begin{cases} class\,0\ if\ x \le threshold \\ class\,1\ if\ x > threshold \end{cases}$

Figure 4.3. Network topology for FOV classification network. Arrows represent layer operations (defined, top right), grey boxes represent feature maps, labelled with number of feature maps (top) and image dimension (left).

4.2.2. Results

Classification metrics for the FOV detection network are reported in Table 4.3. Tile metrics are calculated after applying a binary threshold to the output of the sigmoid layer, which is defined using Youden's index of the tuning set. Entire slice predictions are also reported; these predictions are calculated by averaging the prediction of each tile belonging to the entire slice. This is done to reduce errors arising incorrect prediction due to imaging artifacts, poor probe contact, or off-target scatterers (e.g. clinician touching the probe, mucus). A threshold of 90% of a slice's tiles must be classified into class 1 was manually selected for a slice to be accepted as within the desired FOV. ROC and PRC curves for the test set are presented in Figure 4.4. These curves demonstrate a near perfect classifier, evidenced by the AUC and mAP metrics.

Table 4.3. Classification metrics for FOV detection network, calculated with the test cohort.

<table>
<tr><th>Metric</th><th>Per Tile</th><th>Whole Slice</th></tr>
<tr><td>Threshold (<i>method</i>)</td><td>0.79 (<i>Youden's index</i>)</td><td>0.9 (<i>Manual</i>)</td></tr>
<tr><td>Confusion Matrix</td><td>
<table>
<tr><th></th><th></th><th colspan="2">PREDICTED CLASS</th></tr>
<tr><th></th><th></th><th>CLASS 0</th><th>CLASS 1</th></tr>
<tr><th rowspan="2">ACTUAL CLASS</th><th>CLASS 0</th><td>6497</td><td>121</td></tr>
<tr><th>CLASS 1</th><td>3</td><td>9915</td></tr>
</table>
</td><td>
<table>
<tr><th></th><th></th><th colspan="2">PREDICTED CLASS</th></tr>
<tr><th></th><th></th><th>CLASS 0</th><th>CLASS 1</th></tr>
<tr><th rowspan="2">ACTUAL CLASS</th><th>CLASS 0</th><td>186</td><td>0</td></tr>
<tr><th>CLASS 1</th><td>0</td><td>260</td></tr>
</table>
</td></tr>
<tr><td>Bal. Accuracy</td><td>69.4%</td><td>100.0%</td></tr>
<tr><td>Sensitivity</td><td>99.9%</td><td>100.0%</td></tr>
<tr><td>Specificity</td><td>38.8%</td><td>100.0%</td></tr>
<tr><td>AUC</td><td>1.00</td><td>-</td></tr>
<tr><td>mAP</td><td>1.00</td><td>-</td></tr>
</table>

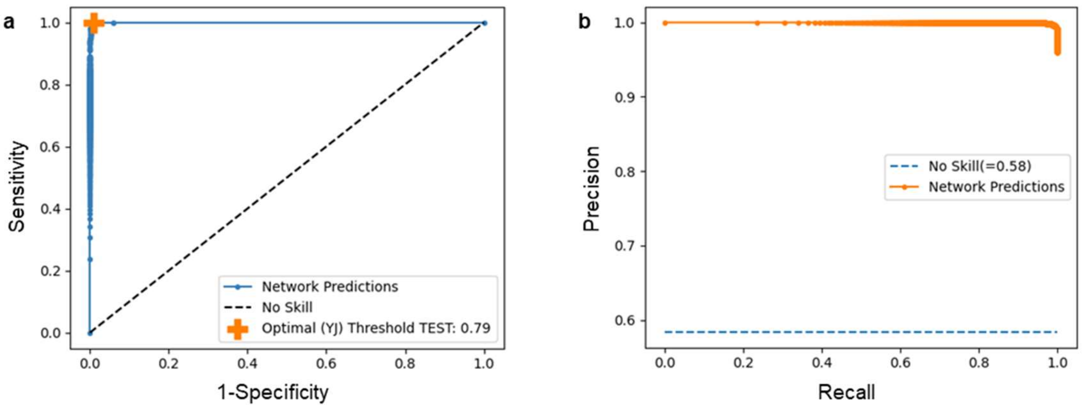


Figure 4.4. FOV tile classification (a) ROC curve; (b) PRC curve. Baseline PRC = 0.58.

4.2.3. Discussion

This network was trained primarily on distinct images containing either good or no probe contact. This bifurcation was selected intentionally to reduce manual labelling, where slices near the tissue-air transition would have required per-tile classification to ensure accurate labels. However, this step of the DL pipeline is designed to separate entire longitudinal slices, and misclassification of individual tiles is managed through post-processing, wherein 90% of tiles must be predicted to have good tissue contact to be processed in further steps.

Network success is evidenced in Figure 4.5, demonstrating invariance to small imaging artifacts (encompassing <30% of the image, panel a), generalizability to tissue morphology, (i.e. unimpacted by non-stratified tissue, panel b), and rejection of off-target scattering materials at a distance from the imaging probe Figure 4.5, panel c).

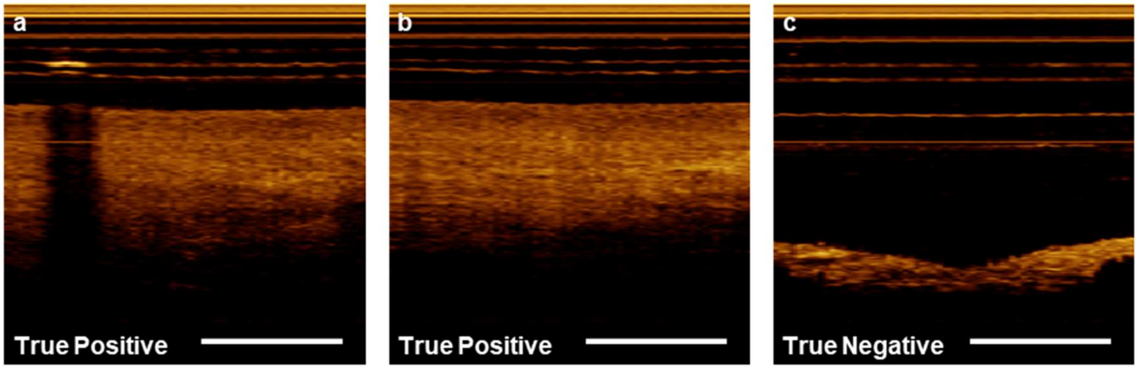


Figure 4.5. **Correct FOV network tile predictions. (a) Correct inclusion of small artifacts; (b) correct inclusion of unstratified tissue; (c) correct exclusion of non oral tissue content. Scale bars 1mm.**

Figure 4.6 details network failures, classified as type 1 (FP, panel a) errors, and type 2 (FN, panels b, c) errors. Type 1 errors arise from off-target scattering materials that present with similar optical properties to oral tissue (suspected mucus) and are close to the imaging probe. Type 2 errors can be attributed to result of large imaging artifacts, obfuscating the tissue (panel b), or poor OCT quality resulting from tissue roll-off at the transition from tissue to air (panel c).

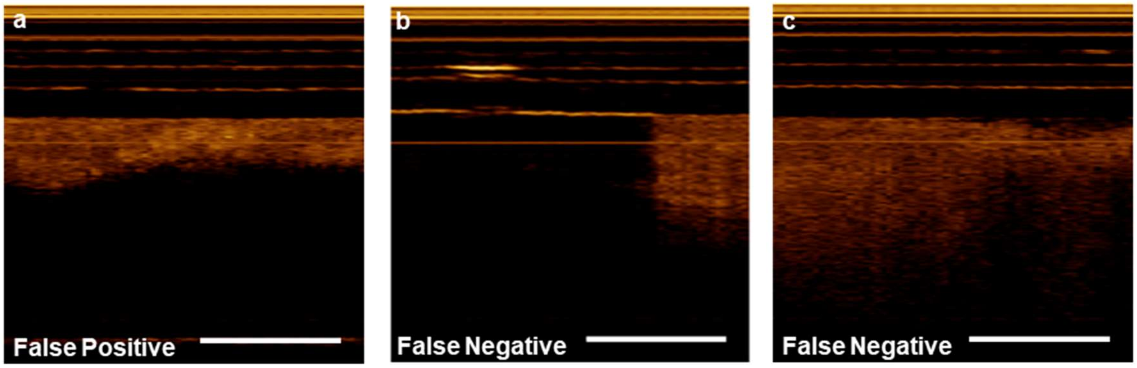


Figure 4.6. **Incorrect FOV network tile predictions. (a) Incorrect exclusion of large artifact; (b,c) incorrect exclusion of tissue. Scale bars 1mm.**

4.3. Epithelial Surface Segmentation

This section presents the segmentation network employed to find the epithelial surface, the top surface of the oral tissue. It can be executed in parallel with the basement membrane segmentation network (Section 4.4) and the artifact detection network (Section 4.5) after the FOV detection model (Section 4.2) has determined the sections from the volume that contain good tissue contact.

Previous attempts at segmenting epithelial and basement membrane surfaces were completed by generating an epithelial layer mask from rater annotations. However, this approach was limited by areas with complete loss of basement membrane; understanding of the cellular composition of dysplastic and malignant tissue informs that for samples with basement membrane destruction, the epithelial layer encompasses the entire imaging depth. Digital representation of this phenomena resulted in large areas of epithelial mask, that once tiled could only be identified through knowledge of neighbouring tiles, rather than local image features. To overcome this issue, separate networks were trained to detect the bottom and top surfaces of the epithelium independently. This section covers the first half of the task: epithelial surface segmentation.

4.3.1. Methods

Dataset Preparation

Epithelial surface reference images were generated using MATLAB to thicken reference annotation lines into segmentation masks. To reduce imbalanced mask content (where only a single pixel is identified per a-line, and high agreement can be achieved by classifying every pixel as background), a line thickening algorithm was applied to the 1D line data. Initially, a coarse mask was generated, thickening the boundary by ± 24 pixels in the a-line direction, for a total mask thickness of 49 pixels. The intention was to use the coarse mask for network pre-training, and thinner (± 12 pixel, ± 4 pixel) masks would be used to fine tune the weights. Magnified sections of longitudinal slices with superimposed boundary masks are detailed in Figure 4.7.

Mask Size	Contralateral example	Pathologic example

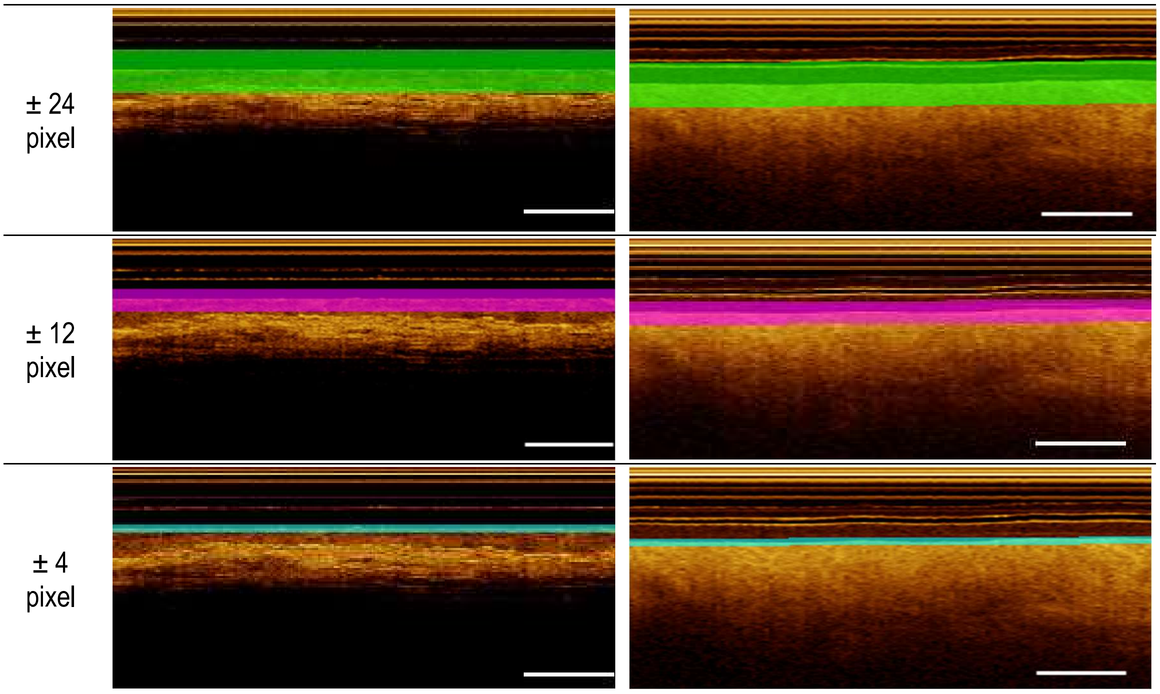


Figure 4.7. Insets of longitudinal slices superimposed with ± 24, ± 12, and ± 4 pixel epithelial surface reference masks, shown on contralateral and pathologic tissue. Scale bars 1mm.

Per the motivations described in Section 4.2.1, masks and corresponding longitudinal slices were divided into 256x256 pixel tiles with a 128 pixel overlap (Figure 4.8), generating 11,356 tiles apiece original slices were 3000-9000x382 pixels.

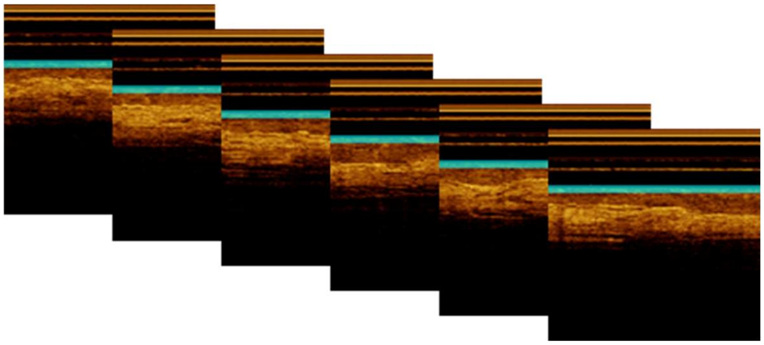


Figure 4.8. Epithelial surface segmentation image tiles with ± 4 pixel reference mask overlay.

A training protocol was defined with the assumption that the imbalanced mask content of the thinner masks would be difficult for the network to learn, where the deeper network layers would lose the spatial context necessary to locate epithelial surface. To evaluate this hypothesis, pre-training was performed with thicker masks, to allow coarse localization of the general region of interest, and subsequently refine predictions through fine-tuning with thinner masks. However, during training it was observed that fine-tuning did not yield significant changes. To validate this observation, metrics were evaluated on networks trained using each mask size.

For network training, tiles were divided into training, tuning and testing cohorts. An ideal split of 70-15-15 percent distribution was identified to increase feature variance in the tuning and testing cohorts, but ensure sufficient data in the training cohort. However, to prohibit data leak between cohorts, slices aquired from the same patient needed to be assigned to the same cohort. The resulting distribution is summarized in Table 4.4. Z-score normalization was applied to all cohorts.

Table 4.4. Distribution of tiles and patients for epithelial segmentation network.

Cohort	No. Tiles (%)	No. Patients
Train	9,073 (73.1)	42
Tune	1,139 (12.6)	8
Test	1,144 (14.1)	9
Total	11,356	59

Network Definition & Parameters

A shallow u-net was used to train the epithelial surface segmentation layer. Network topology was selected by modifying the standard u-net [69], removing the last encoder and first decoder layers, and connecting the bottleneck one step earlier. This adaption was motivated to match the downsampled image size at the bottleneck of the standard u-net, which was developed with input images 512x512 pixels large. Additionally, a shallower network reduces the likelihood of overfitting by decreasing the number of network parameters. Network topology is shown in Figure 4.9; further details on network topology can be found in Appendix B, Table B2. Hyperparameters implemented for network training are summarized in Table 4.1.

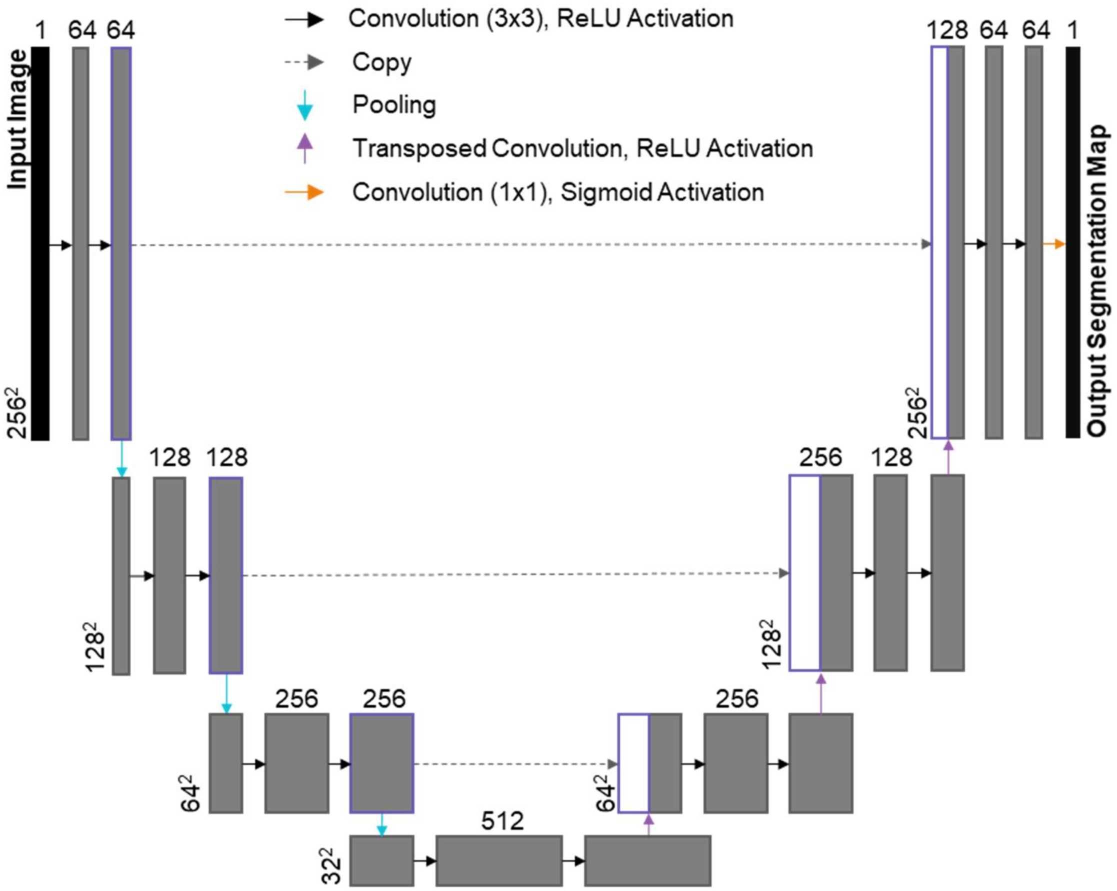


Figure 4.9. Epithelial surface segmentation u-net topology. Arrows represent layer operations, grey boxes represent feature maps, labelled with number of feature maps (top) and image dimension (left). White boxes are copied feature maps of corresponding purple-bound boxes.

Post Processing

The aim of this task was to identify the precise epithelial boundary. This boundary needed to be extracted from the raw output prediction masks, which identified regions of similar thickness to the input pixel mask. Accordingly, a single pixel for each a-line of was identified, indicating the precise depth of the epithelial surface. In the ideal case, this pixel would exist at the middle of the raw prediction mask. However, to reduce the impact of spurious predictions that were occasionally triggered by highly scattering imaging artifacts (Figure 4.10; reference mask in panel a, raw prediction mask in panel b), the median of the first and last identified pixel per a-line was used (panel c). Discussion of boundary selection is further covered in Section 4.3.3.

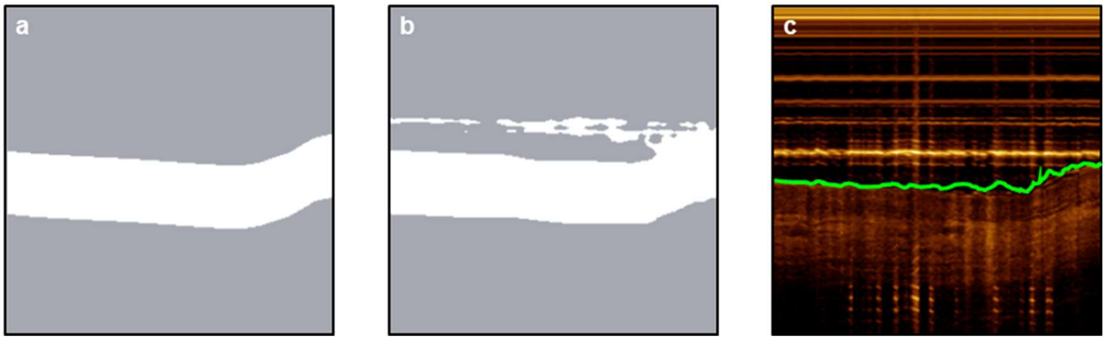
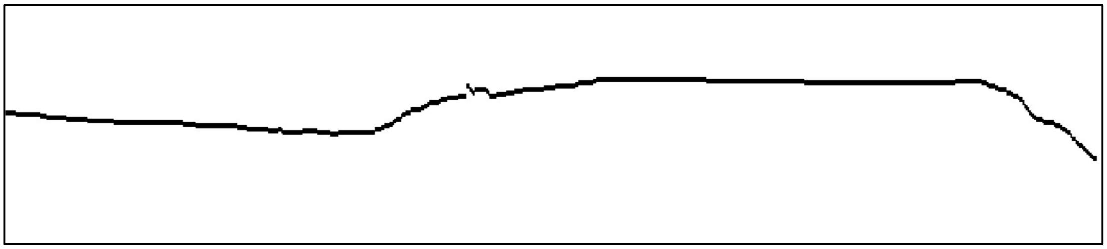
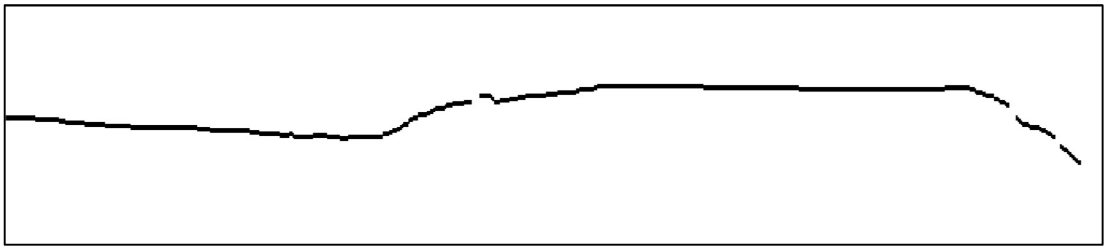
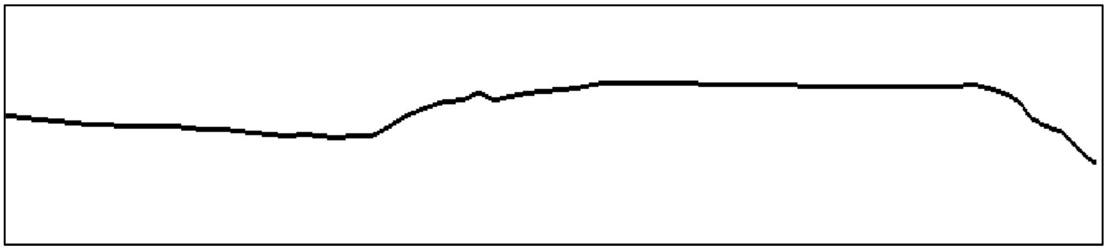
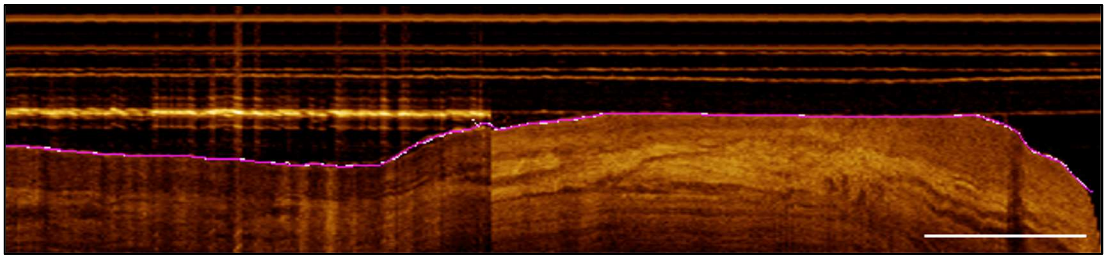


Figure 4.10. Example of prediction error arising from image artifact; (a) ± 24 px reference mask; (b) raw prediction mask; (c) median a-line reference boundary (green) overlay on image tile.

Boundary selection was performed per tile, and all boundaries were stitched to the original slice size. Overlapping regions were averaged, as network predictions of the same information can generally be relied on to be consistent within a few pixels. Subsequently, simple morphological operations were performed to eliminate small, off target regions, and smooth any jagged boundaries that arose during creation of single pixel lines from network prediction masks. Post processing was implemented using MATLAB and was applied to the whole slice after tiles were stitched together. First, disagreements arising from tile overlap were averaged to a single row. Short boundary sections containing less than 5 pixels were subsequently removed. Finally, all boundaries were connected, following the assumption that a surface should exist for the entire longitudinal slice. To smooth the prediction, the entire boundary was resampled at every 10th pixel and reconnected using the pchip algorithm. Morphological operation parameters were selected through empirical evaluation. Steps and sample results are summarized in Table 4.5. Sample results have been slightly thickened for easier visualization.

Table 4.5. Epithelial surface network prediction post-processing pipeline

Step	Description & Example
Input	Median of raw prediction data 

Remove small regions	Objects made up of less than 5 pixels were removed. 
Connect & Smooth	Connected sections less than 5 pixels apart. Downsampled by selecting every 10th pixel along the pullback axis and re-connected using pchip interpolation to smooth edges. 
Compare	Raw (white) and post-processed (magenta) mask superimposed on longitdudinal slice. Scale bar 1mm. 

4.3.2. Results

Table 4.6 presents the segmentation metrics used to evaluate the epithelial surface segmentation network, calculated from the predictions of test cohort on networks trained with ± 24, ± 12, and ± 4 pixel thick masks. As discussed, network fine-tuning with thinner masks did not yield significant improvements, and worse, resulted in increased training time. Both the raw DSC (the average DSC of tiles), and the post-processed DSC (the average DSC of post-processed slices) are reported, as well as the a-line axis pixel error mean and SD (calculated from the post-processed slices). Whereas the DSC is calculated using thickened masks, the mean and SD metrics are calculated from a single pixel thick line. The bias of DSC towards larger mask size is apparent through evaluation of the metrics, with a clear change in DSC while other metrics remain comparable, despite varying training protocols. To ensure fair comparison between the raw and post-processed DSCs, the post-processed prediction has been re-thickened by the same width as the training data. The minimal post processing method meant that negligible improvement was expected between the raw and post processed DSC. Indeed, only the ±12 pixel training protocol demonstrated any improvements, with other protocols demonstrating a slightly decreased DSC; this slight loss is permissible given the limitations of the DSC metric, discussed prior. As post-processing steps ensure the presence of a mask for the entire slice, PA =100%, and NA = 0%.

Table 4.6. Epithelial surface segmentation metrics, calculated with the test cohort.

Protocol	DSC_{raw}	DSC_{post}	μ ± σ
± 24 pixel	98.9	98.8	0.53 ± 0.96
± 12 pixel	97.6	97.8	0.52 ± 0.87
± 4 pixel	94.3	94.1	0.53 ± 1.01

With the metrics being nearly equivalent, and acknowledging the bias of the DSC towards larger objects, the smallest pixel error mean and variance was used to select the network trained with ± 12 pixel thick masks as the best. This selection is further supported through analysis of the histogram of pixel errors, seen in Figure 4.11, with the 12-pixel mask featuring the smallest amount of spurious, off target predictions.

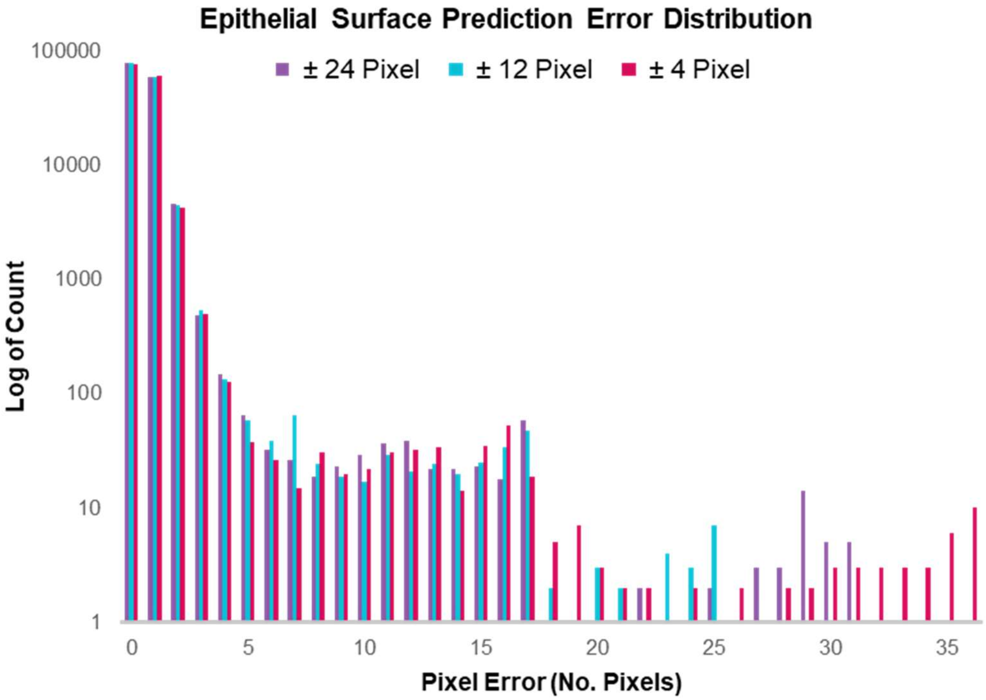


Figure 4.11. Histogram of epithelial surface segmentation a-line depth error for u-net trained on ± 24 pixel masks (purple); ± 12 pixel masks (blue); and ± 4 pixel masks (pink).

Training and tuning curves summarizing the DSC (Figure 4.12), loss, (Figure 4.13) and LR (Figure 4.14) at each epoch during training using the ± 12 pixel training protocol are shown below. The early stopping cut-off point is also included in each figure.

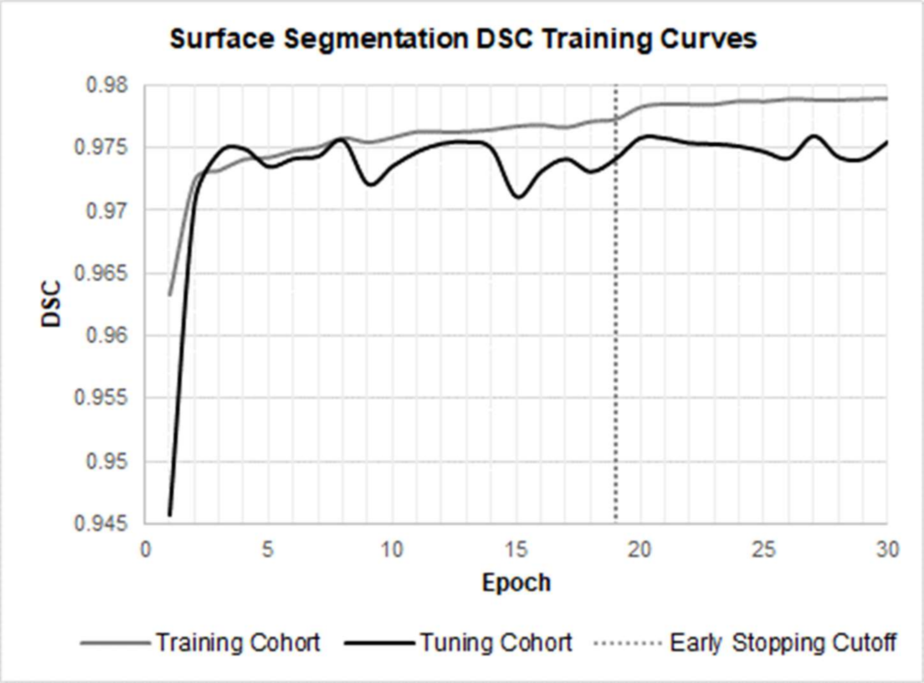


Figure 4.12. DSC training and tuning curves for epithelial surface segmentation network training.

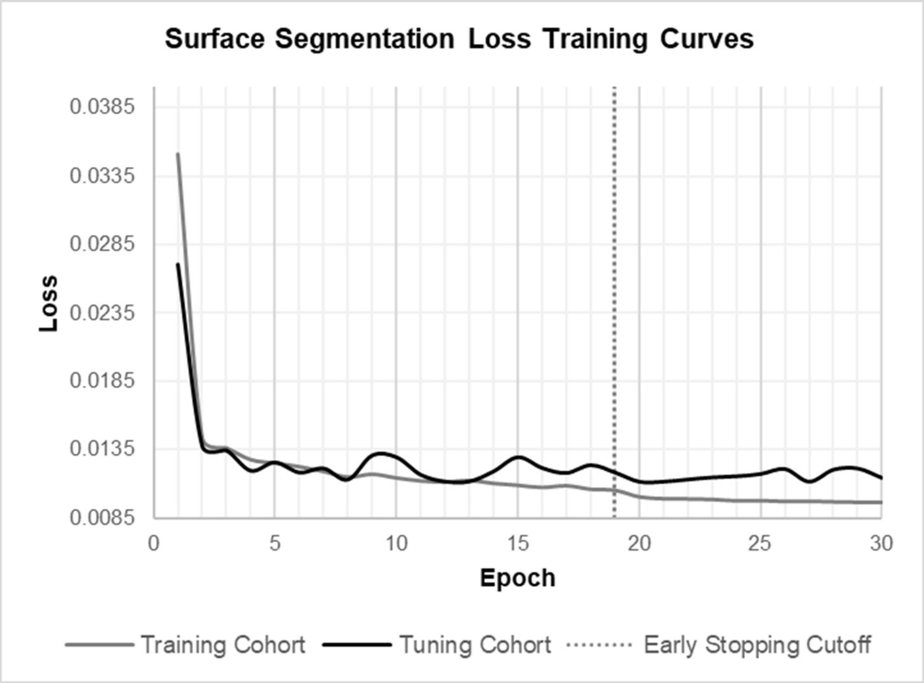


Figure 4.13. Loss training and tuning curves for epithelial surface segmentation network training.

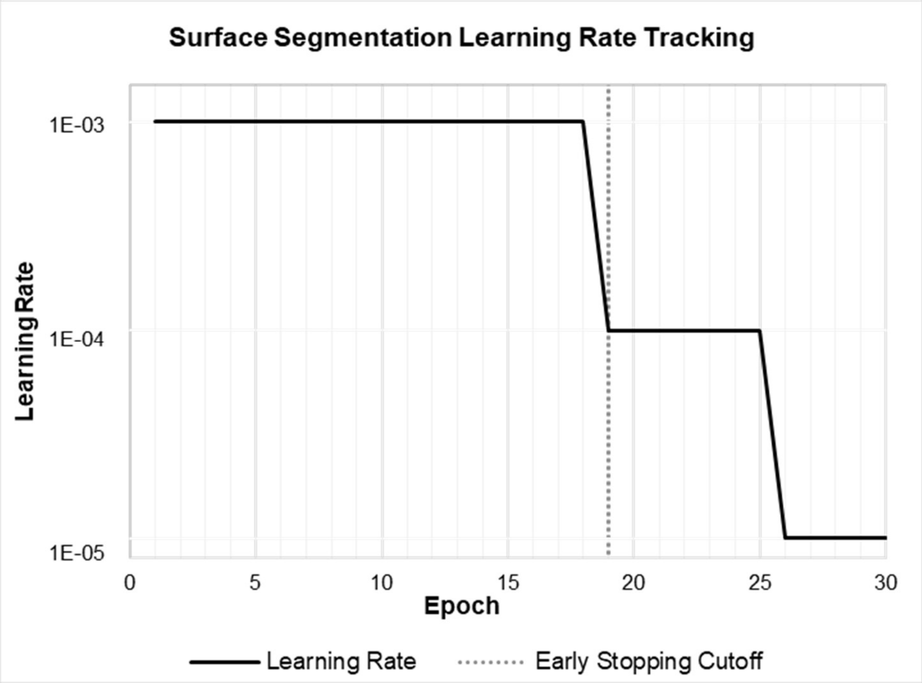


Figure 4.14. LR adaptions for epithelial surface segmentation network training.

4.3.3. Discussion

Selection of the Best Network

Networks were trained on thickened masks to reduce class imbalance; it is not unexpected that the ± 12 pixel training protocol generated the best results, balancing the necessary precision to avoid basing predictions on the protective sheath artifacts, but providing sufficient spatial information that the references were useable at the lowest levels of the network. In fact, it is likely that the networks trained on the ± 4 pixel mask lost important image context at the lowest levels of the network, where the image has been downsampled by a factor of 8 and the region of interest identified in the reference has been downsampled to a single pixel.

Generating a Boundary Detection from Raw Network Predictions

The raw network predictions identified a region of interest of similar size to the reference masks. The epithelial surface boundary could to be extracted from this prediction by selecting median value of the raw mask limits for each a-line. Alternatively, the mean of these values could have been selected, which would be equivalent if the network was guaranteed to only identify one region per a-line. However, as shown in

Figure 4.10 panel b, confounding features (e.g. highly scattering imaging artifacts) can trigger spurious predictions and result in discontinuous intra a-line regions. Taking the mean of the limits was more likely to result in off-target boundary. If further precision was deemed necessary, methods could include selection of the mean of the largest identified region, or using classical processing methods to detect the correct edge within the identified region; these methods require more extensive processing and computational time. For the purposes of this research, the less precise median of the limits was deemed sufficient, but applications of this method in other datasets may require further exploration.

Demonstrating Network Success

While the segmentation metrics used to quantify the network success are bounded, with a DSC of 1 indicating perfect agreement, and a mean and SD of 0 also indicating perfect agreement, achieving perfect agreement is not a reasonable expectation for deep learning segmentation tasks. Instead, network success is demonstrated by comparing the reference-prediction metrics to inter-rater metrics. Segmentation is a subjective task, resulting in human raters being prone to error or differences of opinion regarding boundary location. As such, a network can be concluded successful if the reference-prediction metrics are within the distribution of inter-rater metrics.

Results for rater-prediction, rater-reference and inter-rater calculations of the DSC, and mean and SD are shown in Table 4.7 and Table 4.8, respectively. Network prediction metrics are calculated from the test cohort evaluated with the ± 12 pixel training protocol. Raters that did not share longitudinal slices are marked by N/A.

Table 4.7. Rater-prediction, rater-reference and inter-rater DSCs for epithelial surface segmentation network. N/A indicates raters that did not intersect annotations.

DSC	Reference	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Prediction	97.8	96.2	96.5	96.9	97.4	96.1
Reference		96.2	96.4	96.5	97.2	95.2
Rater 1			95.8	95.9	96.8	96.2
Rater 2				96.0	95.8	N/A

Rater 3					96.8	95.5
Rater 4						N/A

Table 4.8. Rater-prediction, rater-reference and inter-rater a-line depth pixel error mean and SD metrics for epithelial surface segmentation network. N/A indicates raters that did not intersect annotations.

μ ± σ	Reference	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Prediction	0.52 ± 0.88	0.88 ± 1.05	0.75 ± 1.14	0.71 ± 0.89	0.63 ± 0.69	0.95 ± 0.84
Reference		0.89 ± 0.89	0.80 ± 0.81	0.86 ± 0.76	0.69 ± 0.90	1.20 ± 0.96
Rater 1			0.96 ± 0.98	0.96 ± 0.88	0.79 ± 0.73	0.87 ± 0.89
Rater 2				0.88 ± 0.81	0.89 ± 1.03	N/A
Rater 3					0.78 ± 0.77	1.02 ± 1.00
Rater 4						N/A

Comparison of the above results to the epithelial surface u-net reference-prediction metrics (DSC = 97.8, μ = 0.52, σ = 0.88) demonstrates that DL methods offer an improvement over manual annotation metrics, outperforming all inter-rater metrics. While poor inter-rater agreement may be attributed in part to labelling errors rather than misidentification of the surface of interest, this type of error supports the motivation for automated methods. It is worth noting that in physical space, a single pixel error represents a 10μm error, demonstrating strong model performance.

Challenges & Limitations

Challenges identified in this task included a small selection (n=3) of longitudinal pullbacks where unknown materials occluded the image, as seen in Figure 4.15. This substance is hypothesized to be mucous or saliva. During annotation, raters were asked to draw boundaries against the true tissue surface and exclude this matter. However, due to low number of samples, the network was not able to accurately learn this boundary, and commonly created incorrect, but conceivable annotations. Examples of correct predictions are marked by a white arrow, while incorrect predictions are marked with a white X.

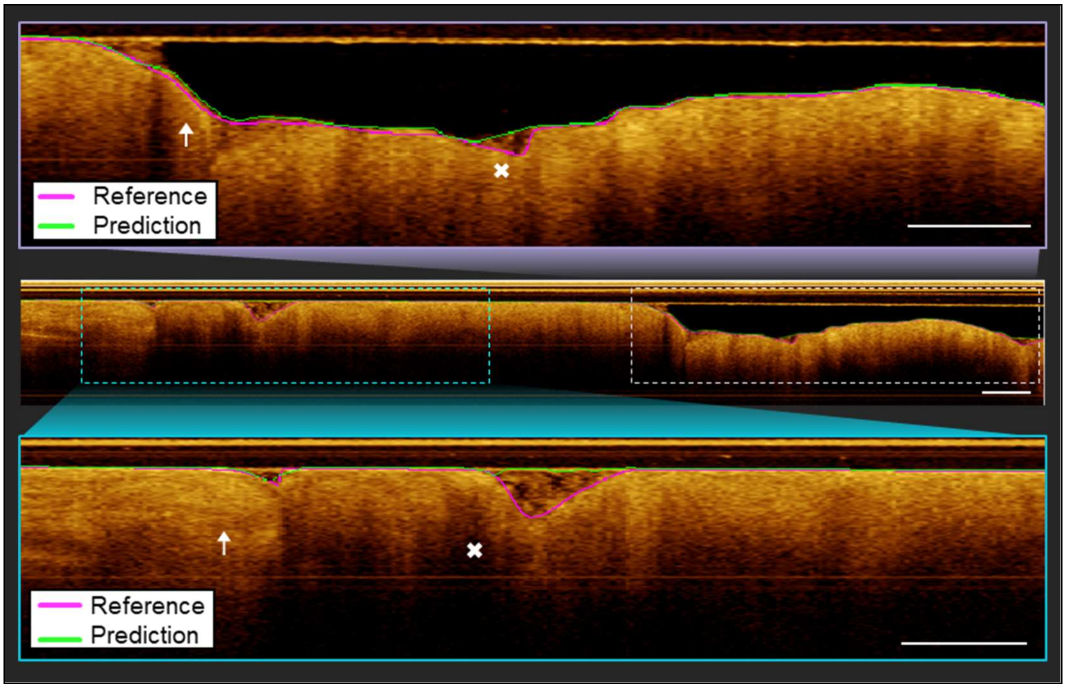


Figure 4.15. Examples of epithelial surface segmentation error due to unknown surface occlusions. Incorrect predictions are marked with a white x and correct predictions are marked with a white arrow. Scale bars 1mm.

Another limitation of this method is recognizing folds that occur in softer tissues of the oral mucosa, e.g. the tongue and buccal mucosa. Again, presentation of these samples is limited, with only 2 cases across the entire cohort. Accordingly, the network does not provide accurate predictions, as shown in Figure 4.16. This limitation is noted but efforts have not been made to correct this error due to limited sample size. Notably, should sufficient samples exist to accommodate these cases, current post-processing methods would not retain the information, and would need to be adjusted accordingly.

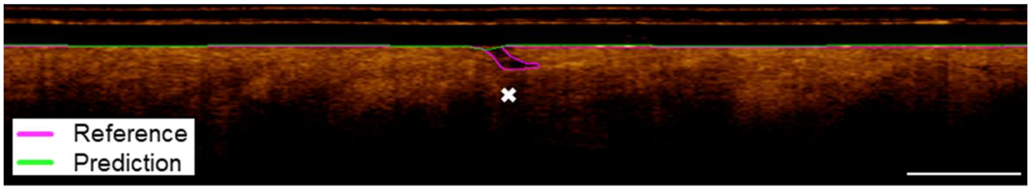


Figure 4.16. Example of epithelial surface segmentation error due to tissue folding or hole.

4.4. Basement Membrane Segmentation

This section presents the network developed to detect acellular basement membrane, a thin layer that demarcates the epithelial and stromal layers.

The basement membrane segmentation task followed a similar method to the epithelial surface task. In contrast to the surface segmentation task however, it is no longer expected that a continuous boundary exists, as the morphological changes that arise in oral cancer cases result in a destruction of this boundary. Thus, the primary goal of this task is to not only identify regions where the basement membrane exists (allowing measurement of epithelial thickness), but of equal importance, locate regions where this transition zone cannot be visualized.

4.4.1. Methods

Dataset Preparation

Basement membrane reference images were generated using MATLAB to thicken rater annotation lines into segmentation masks. Per motivations described in section 4.3.1, three segmentation masks were developed per image. Healthy appearing (contralateral) slices were expected to contain continuous segmentations across the basement membrane, while pathologic slices may contain regions where the basement membrane could not be visualized; these breaks were left blank. Magnified longitudinal sections with superimposed boundary masks for contralateral and pathologic examples are detailed in Figure 4.17.

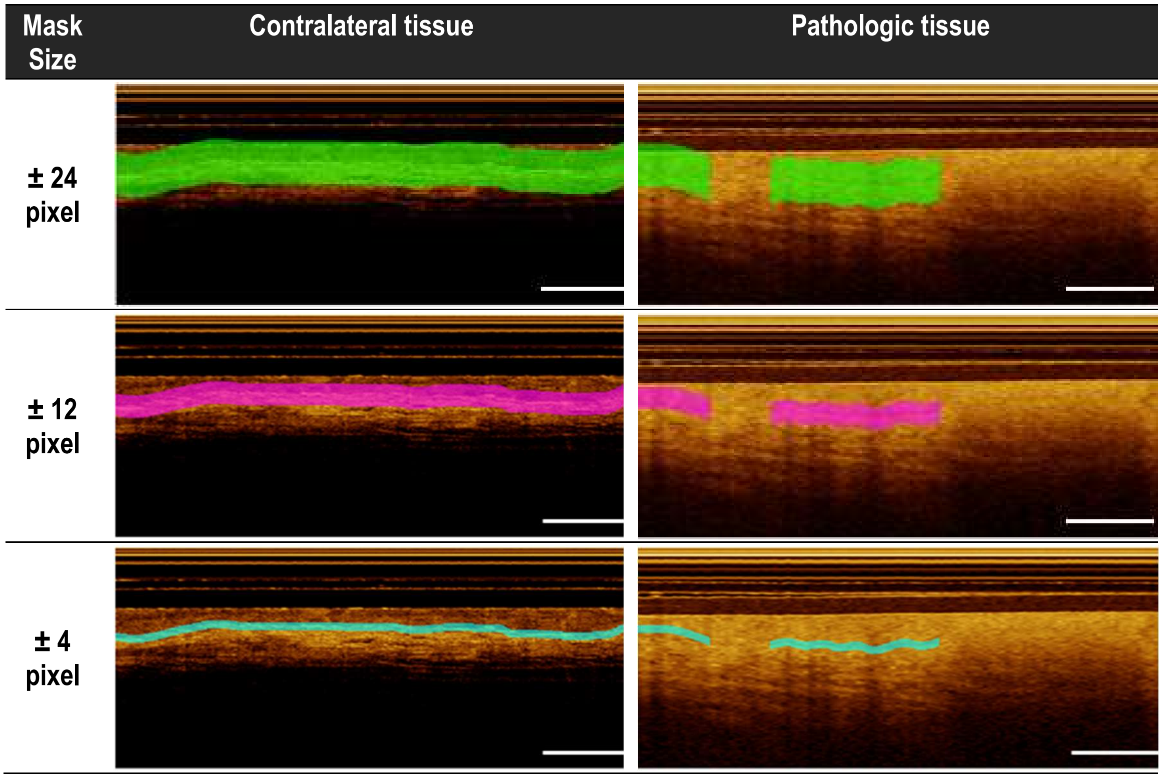


Figure 4.17. Insets of longitudinal slices superimposed with ± 24, ± 12, and ± 4 pixel basement membrane reference masks, shown on contralateral and pathologic tissue. Scale bars 1mm.

Various pre-training and fine-tuning protocols were evaluated during the model selection process, exploring the same concepts discussed in Section 4.3.1. However, in this task it was observed that the fine-tuning approach did improve predictive power. This will be further discussed in Sections 4.4.2 and 4.4.3.

Once again, each mask type and corresponding longitudinal slice was divided into 256x256 pixel tiles, with 128 pixel overlap, generating 11,356 tiles apiece. Sample tiles are shown in Figure 4.18.

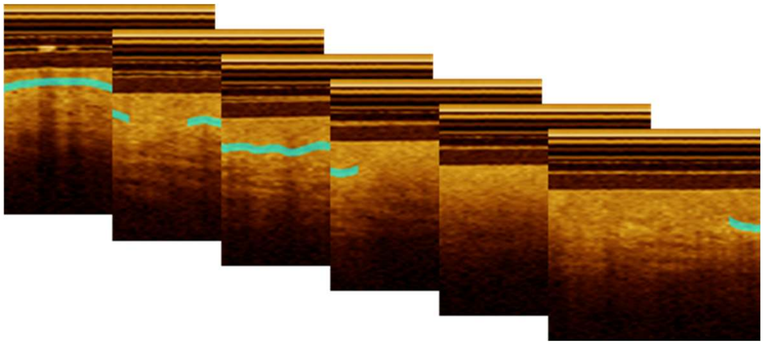


Figure 4.18. Image tiles with basement membrane reference annotation overlay

Tiles were divided into training, tuning and testing cohorts, again with an ideal distribution of 70-15-15 percent. In addition to ensuring that all patients would be assigned to the same cohort, stratified sampling was applied to ensure an even distrubution of tiles with different states of basement membrane continunity, subcategorized as complete, broken or partial and missing. This distribution is summarized in Table 4.9. Z-score normalization was applied to all cohorts.

Table 4.9. Cohort Distribution of basement membrane tiles, patient count and tile features.

Cohort	No. Tiles (% of cohort)			No. Patients
	Complete	Broken	Missing	
Train	6,597 (72.7)	1,141 (12.6)	1,335 (14.7)	42
Tune	842 (73.9)	141 (12.4)	156 (13.7)	8
Test	810 (70.8)	159 (13.9)	175 (15.3)	9
Total	8,294	1,441	1,666	59

Network Definition & Parameters

A shallow u-net was modified to train the basement membrane segmentation network. The u-net (topology shown in Figure 4.19), was adjusted to provide the output prediction of the previous epoch onto the input of the proceeding epoch, providing improved localization of the area of interest. The first pass of the network received blank tiles in the secondary input. Other modifications were made using the same methods described in Section 4.3.1. Further details on network layers can be found in Appendix B, Table B3. Hyperparameters implemented for network training are summarized in Table 4.1.

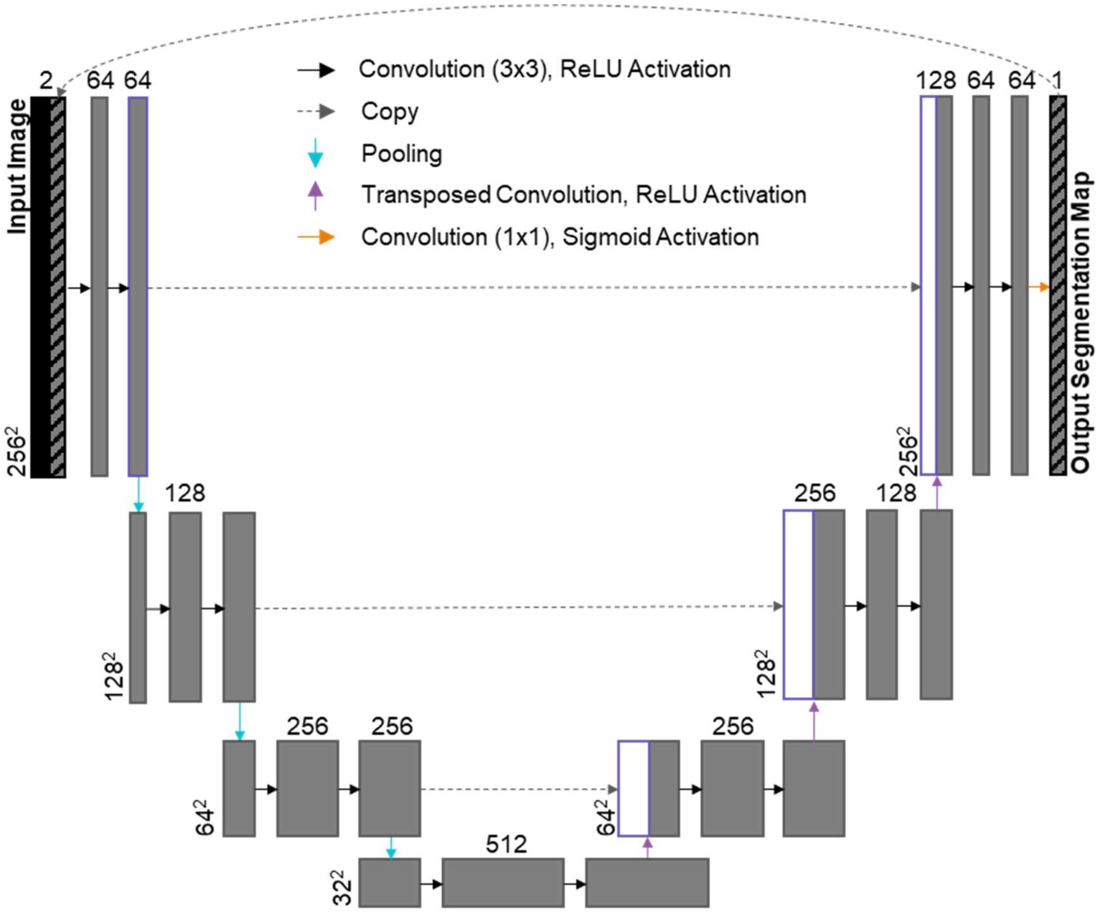
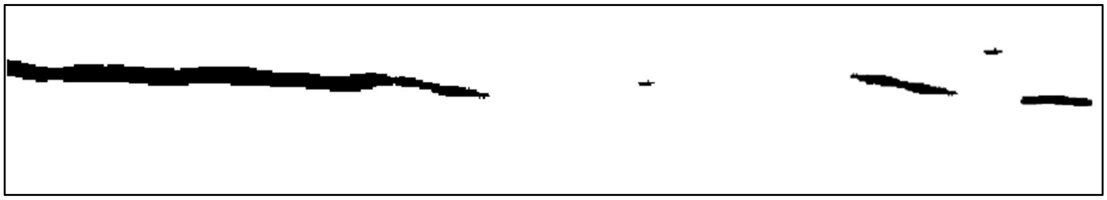
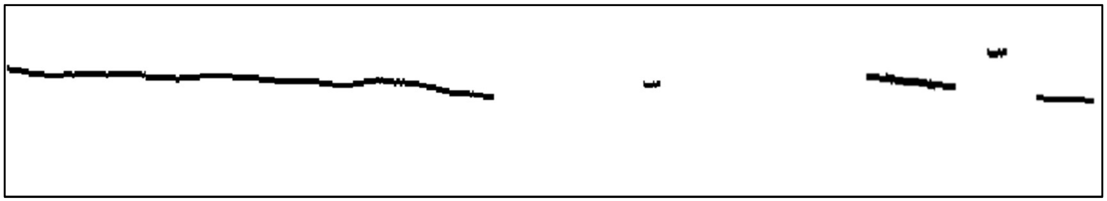
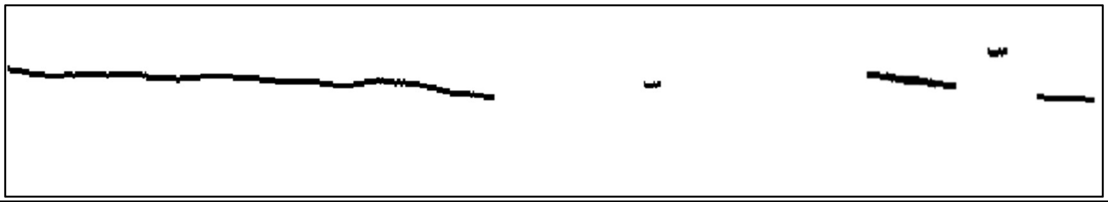
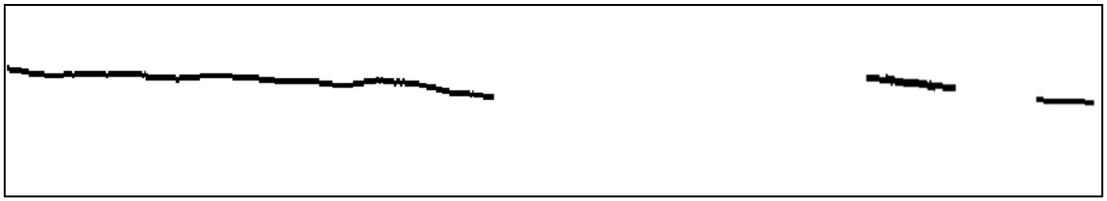
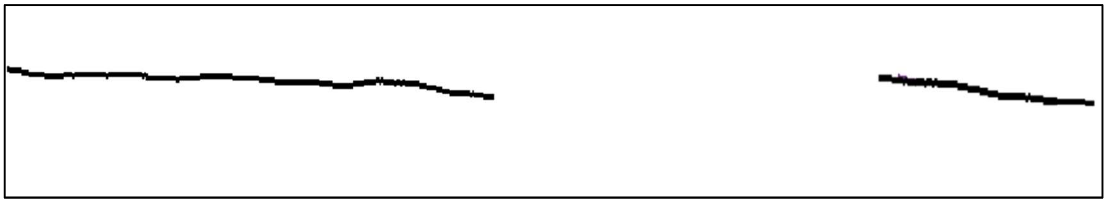


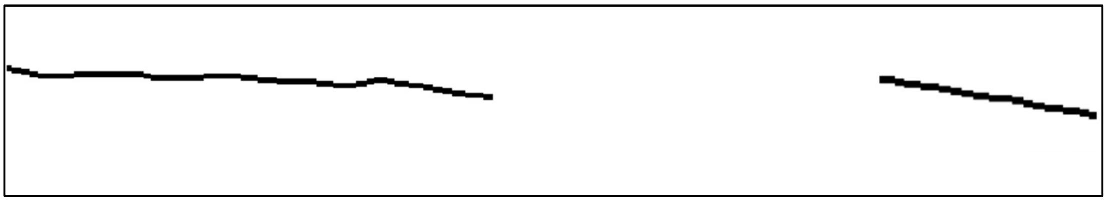
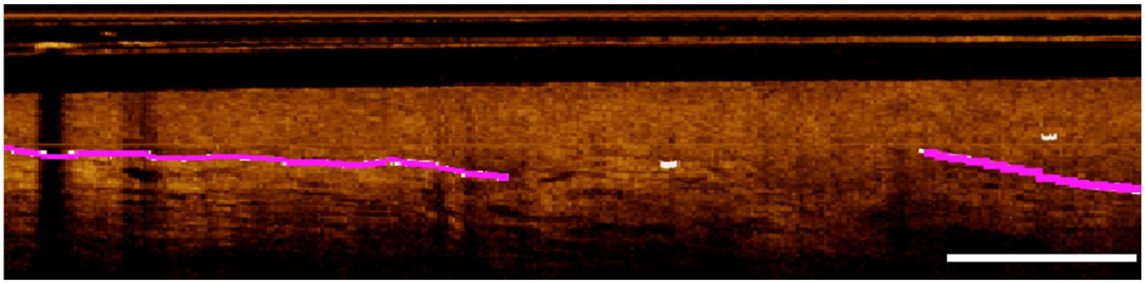
Figure 4.19. Basement membrane u-net topology. Arrows represent layer operations, grey boxes represent feature maps, labelled with number of feature maps (top) and image dimension (left). White boxes are copied feature maps of corresponding purple-bounded box.

Post Processing

Post processing for the basement membrane segmentation task was more extensive than the simpler epithelial surface segmentation task. Post processing was applied to the entire slice, which required stitching raw output tiles; overlapping regions were handled by including all identified pixels within the overlap. Demonstrated graphically in Table 4.10, steps included generating a boundary through selection of the median tile, connecting small gaps (<10 pixels along the pullback axis) through linear interpolation, removing small predictions (< 25 pixels), connecting small gaps (< 30 pixels along the pullback axis) through linear interpolation, and smoothing predictions through downsampling and interpolation with the p-chip algorithm. For linear interpolation steps, predictions that were within the assigned pullback-axis limits but were more than 10-pixels apart along the a-line direction were not connected, as this did not represent the biological organization of oral tissue.

Table 4.10. Basement membrane network prediction post-processing pipeline

Step	Description
Input	Stitched raw prediction data 
Generate Boundary	Median of raw prediction data 
Link proximal boundaries	Boundaries that contained breaks of less than 10 pixels along the pullback axis were linked using linear interpolation (not pictured in this example). 
Remove small regions	Objects made up of less than 25 pixels were removed. This parameter was selected with the knowledge that no reference boundaries were less than 25 pixels. 
Link proximal boundaries	Boundaries that contained breaks of less than 25 pixels along the pullback axis were linked using linear interpolation 

Smooth	Downsampling by selecting every 10th pixel along the pullback axis and re-connected using pchip interpolation to smooth edges.
	
Compare	Raw (white) and post-processed (magenta) mask superimposed on slice. Scale bar 1mm.
	

4.4.2. Results

Segmentation metrics for the basement membrane segmentation test set are presented in Table 4.11. As with the surface segmentation metrics, both the raw DSC (the average DSC of tiles), and the post-processed DSC (the average DSC of post-processed slices) are reported, as well as the a-line depth pixel error mean and SD (calculated from the post-processed slices). The change in DSC (Δ DSC) after post-processing is also shown. PA and NA are reported to quantify agreement regarding agreement between presence and lack of basement membrane prediction; these are also calculated on the post-processed predictions. Several combinations of fine-tuning models are evaluated, listed from least to most complex.

Table 4.11. Basement membrane segmentation metrics, calculated with the test cohort. Best results are bolded.

Model	Training Protocol	DSC_{raw}	DSC_{post}	Δ DSC	PA (%)	NA (%)	μ ± σ
A	± 12 pixel	**86.1**	**91.2**	5.1	95.8	80.9	1.21 ± 1.81
B	± 4 pixel	77.1	75.5	-1.5	93.1	75.1	**0.96 ± 1.31**
C	± 12 → ± 4 pixel	77.3	82.9	**5.6**	**95.8**	**81.7**	1.15 ± 1.55
D	± 24 → ± 4 pixel	77.1	76.0	-1.1	95.1	80.1	1.09 ± 1.98
E	± 24 → ± 12 → ± 4 pixel	75.9	78.6	2.7	95.3	81.2	1.20 ± 1.58

The affinity of the DSC towards larger mask objects can once again be observed, as model A generates the largest masks, and thus cannot be fairly compared to the other protocols. Additionally, Model A performed more poorly than other models in other categories. While model B presents superior mean and SD metrics, this simpler model suffers from the lower PA and NA values, indicating poor representation of basement membrane detection. Model C yielded the highest DSC among comparable models and yielded the highest PA and NA values, indicating the best agreement with regions both with and without an identifiable visible basement membrane. Models D and E present similar metrics to model C, but with slightly worse results. Model E also requires an undesirable longer training time. Thus, model C was selected as the best for this application. The distribution of a-line depth error for models B, C and D are shown in Figure 4.20. For simplicity, only these models are presented. The confusion matrix used to calculate PA and NA is also presented in Figure 4.21

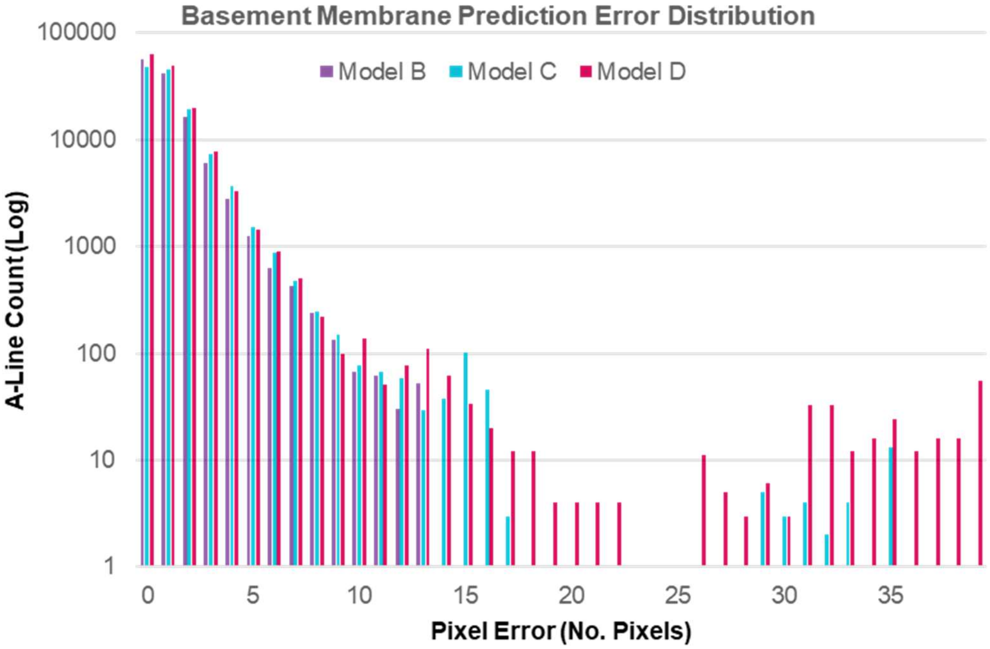
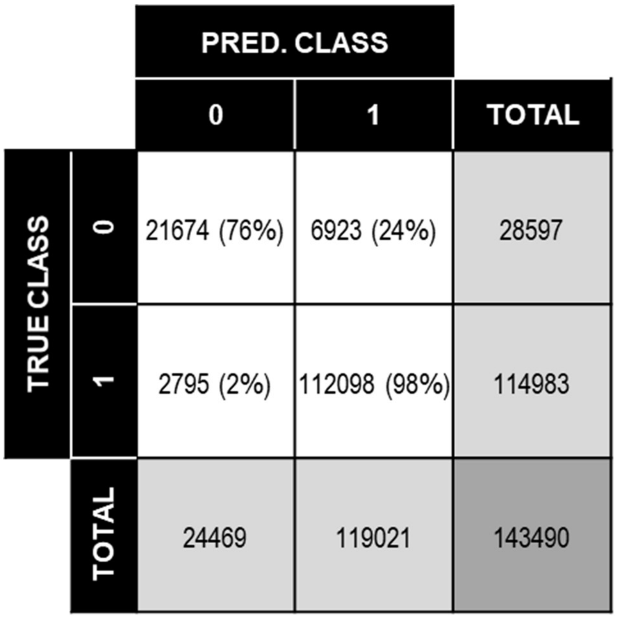


Figure 4.20. Histogram of basement membrane segmentation absolute a-line depth error for u-net model B (± 4px masks, purple); model C (pre-train ± 12px, fine-tune ± 4px masks, blue); and model D (pre-train ± 24px, fine-tune ± 4px masks, pink).



		PRED. CLASS		
		0	1	TOTAL
TRUE CLASS	0	21674 (76%)	6923 (24%)	28597
	1	2795 (2%)	112098 (98%)	114983
	TOTAL	24469	119021	143490

Figure 4.21. Confusion matrix for model C predictions (pre-train ± 12px, fine-tune ± 4px masks). Percent indicates proportion of the true class captured.

Training and tuning curves summarizing the DSC (Figure 4.22), loss (Figure 4.23) and LR (Figure 4.24) at each epoch for model C (pre-training with ± 12 pixel masks, fine-tuning with ± 4 pixel masks) are shown below, as well as the early stopping cut-off point. The expected decrease in the DSC curve can be observed, but the concurrent decrease in loss indicates continued improvement.

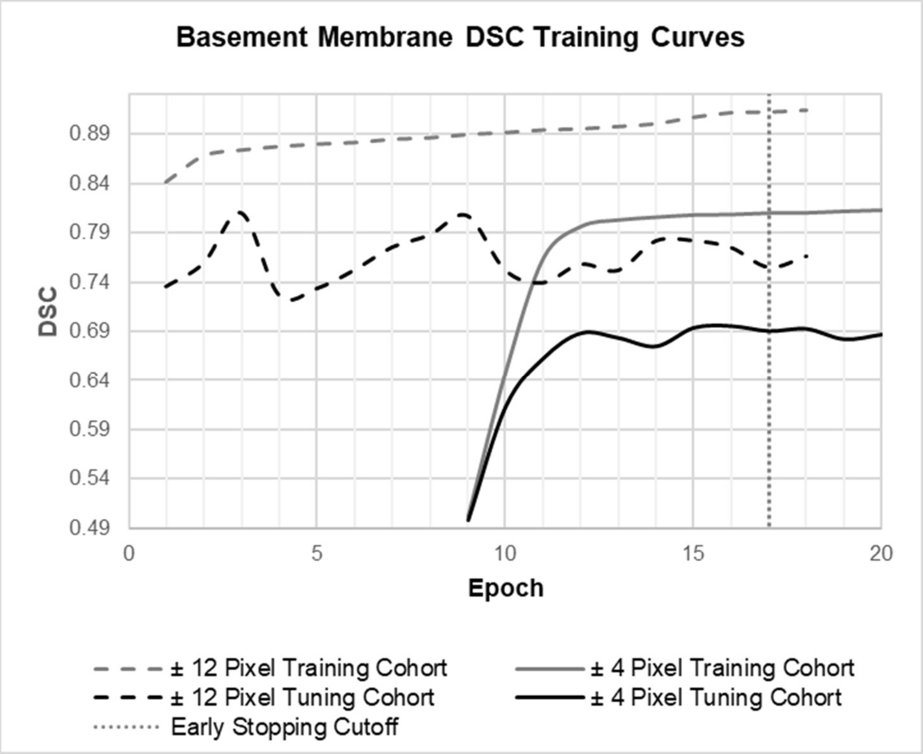


Figure 4.22. **DSC training and tuning curves for basement membrane segmentation network training (u-net model C).**

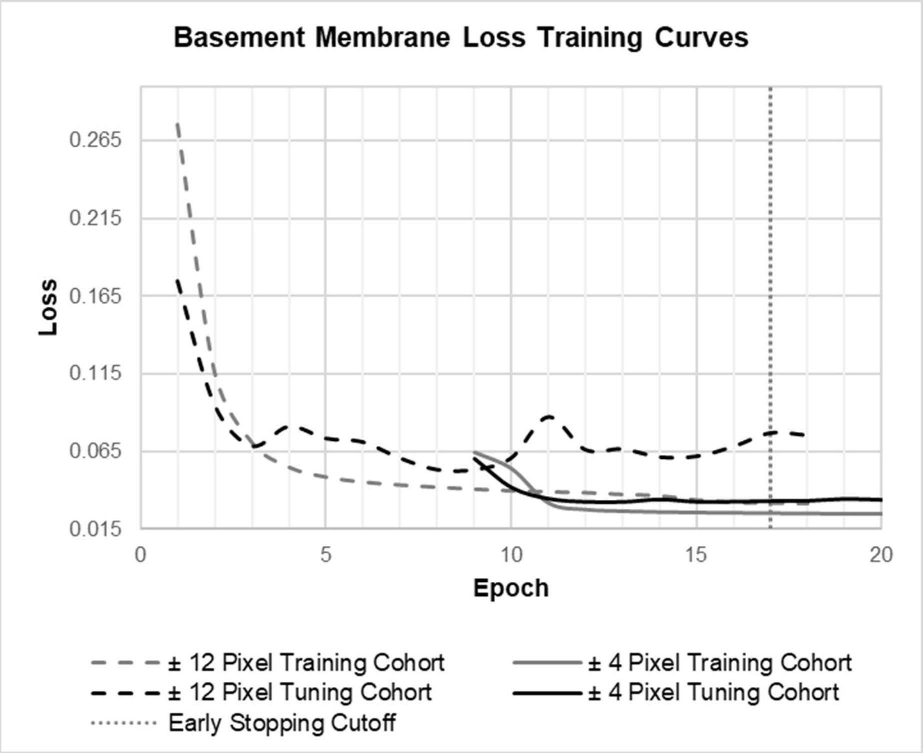


Figure 4.23. **Loss training and tuning curves for basement membrane segmentation network training (u-net model C).**

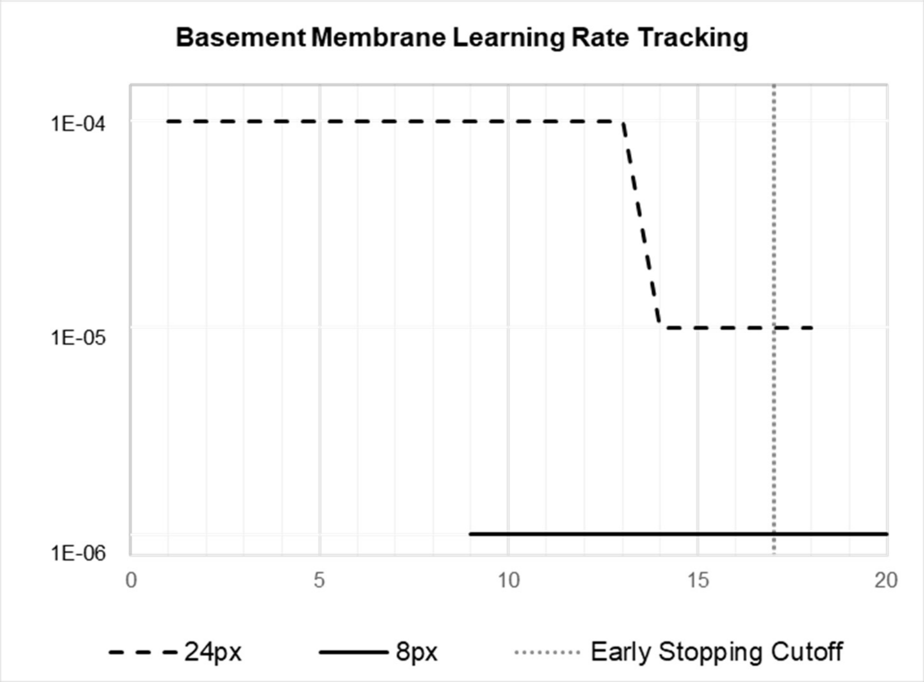


Figure 4.24. Adaptive LR scheme for basement membrane segmentation network training (u-net model C).

4.4.3. Discussion

A Note about the DSC

As discussed in Section 2.4.2, while the DSC is a common metric applied for segmentation tasks, small objects such as the boundaries presented in this thesis may present drastically different results when there is only slight disagreement. The application of post-processing methods presents a second limitation of the DSC. While the raw DSC metrics are calculated tile-wise, the post-processed DSC scores are reported for the entire tile. Demonstrated graphically in Figure 4.25, if an incorrectly prediction is only identified in one out of five tiles, then the average DSC is 80. However, once all tiles have been stitched together this erroneous prediction dominates the entire prediction resulting in a DSC of 0. This condition likely accounts for the negative Δ DSC observed in Models B and D.

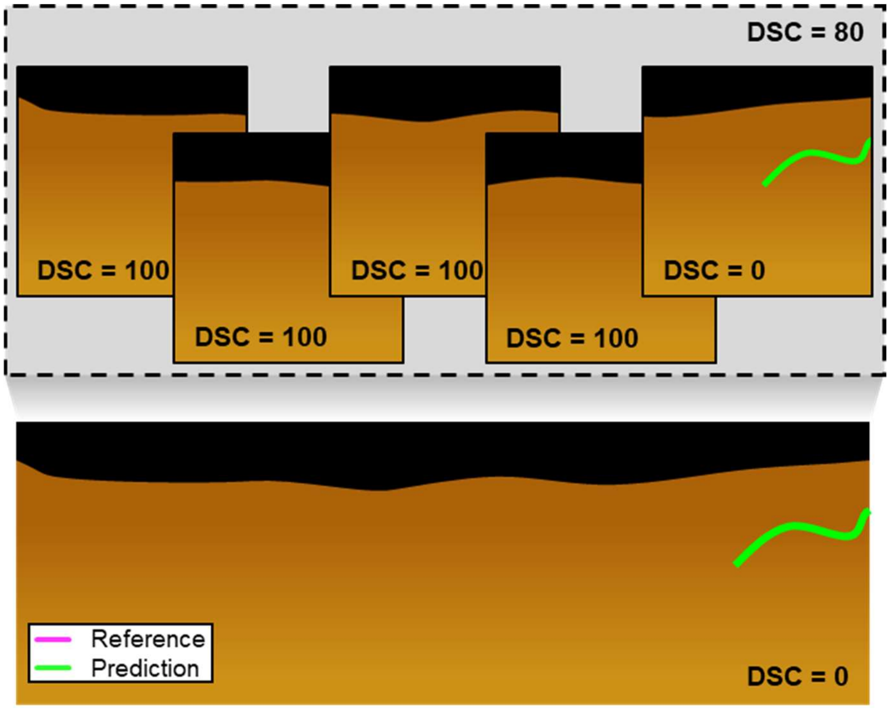


Figure 4.25. Comparison of tile DSC averaging vs DSC of entire slice.

The DSC is reported to better align with best practices for medical segmentation tasks, but should not be treated as a standalone metric to measure model success, but should instead be used in conjunction with the PA, NA and pixel error measurements, as reported above.

Selection of the Best Network

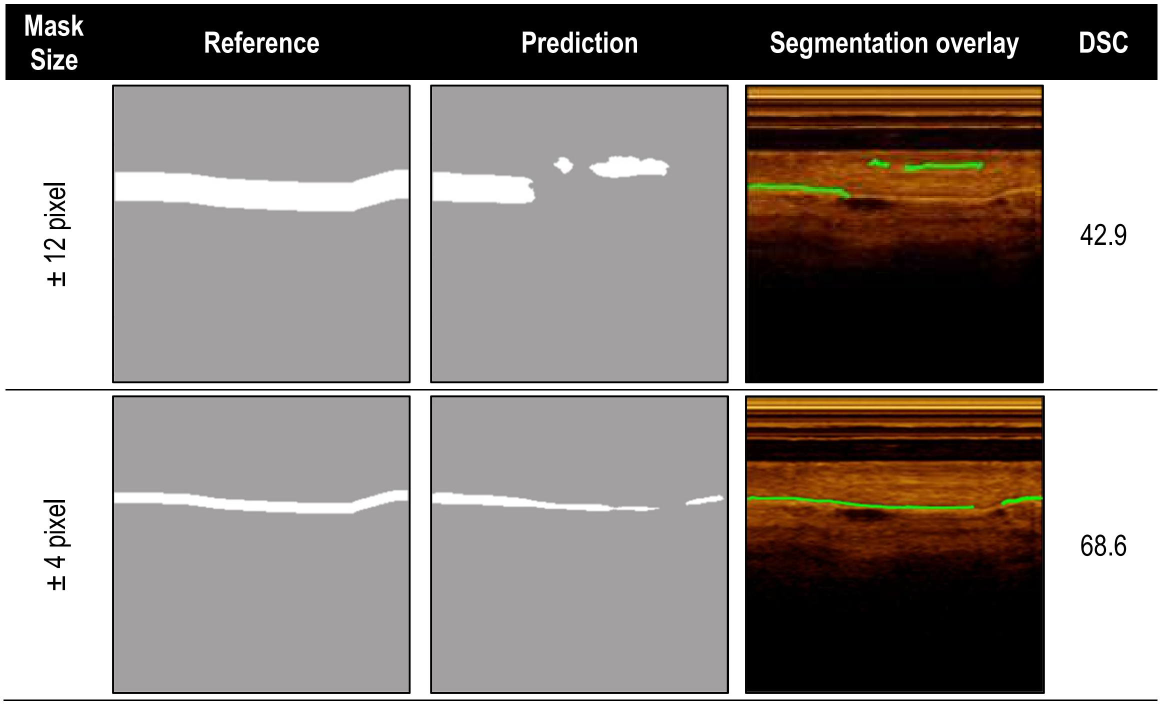
Training of the simpler task of epithelial surface segmentation provided insight that pre-training ± 24 pixel masks did not yield improvements over pre-training with ± 12 pixel masks. However, contrary to the surface segmentation task, fine-tuning with thinner masks did generate meaningful improvements. This difference may be attributed to the more challenging task, or the adaption of the training protocol, where the input layer contained both the input image tile, and the prediction of the previous epoch. This model also had the best response to post-processing methods, despite the algorithm being developed using the raw predictions of model E (± 24 → ± 12 → ± 4 pixel mask training protocol).

Validating Post Processing Methods

To explore the effect of network fine-tuning and post-processing methods, two scenarios are presented below. The first illustrates the case where fine tuning resulted in an improved DSC, and the second, a case where DSC decreased. In both scenarios, progression of predictions from the ± 12 pixel mask pre-training to the ± 4 pixel mask fine-tuning is displayed, along with the results of post-processing algorithm. As the scenarios explore the progression from raw predictions to post-processed predictions, only the DSC is used to quantify differences.

Scenario 1: Increased DSC after fine-tuning

In this scenario, improved localization of the basement membrane resulted in an increased DSC. Summarized Figure 4.26, undesirable gaps are present in the raw prediction of the fine-tuned network, but application of post-processing techniques generates a reasonable segmentation.



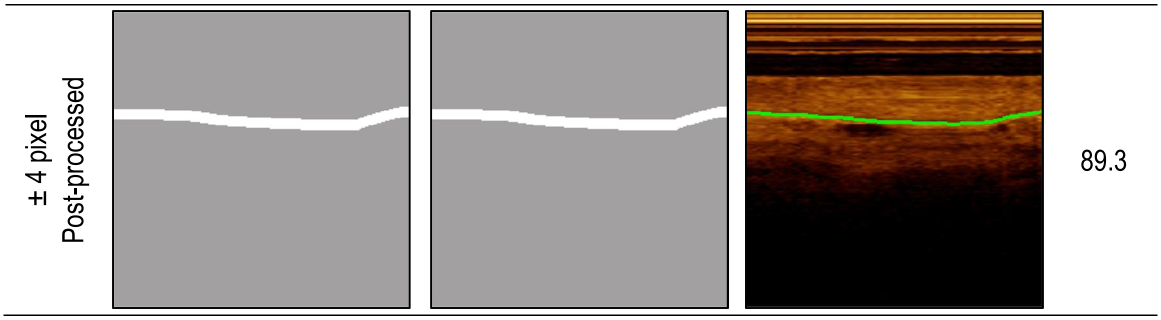
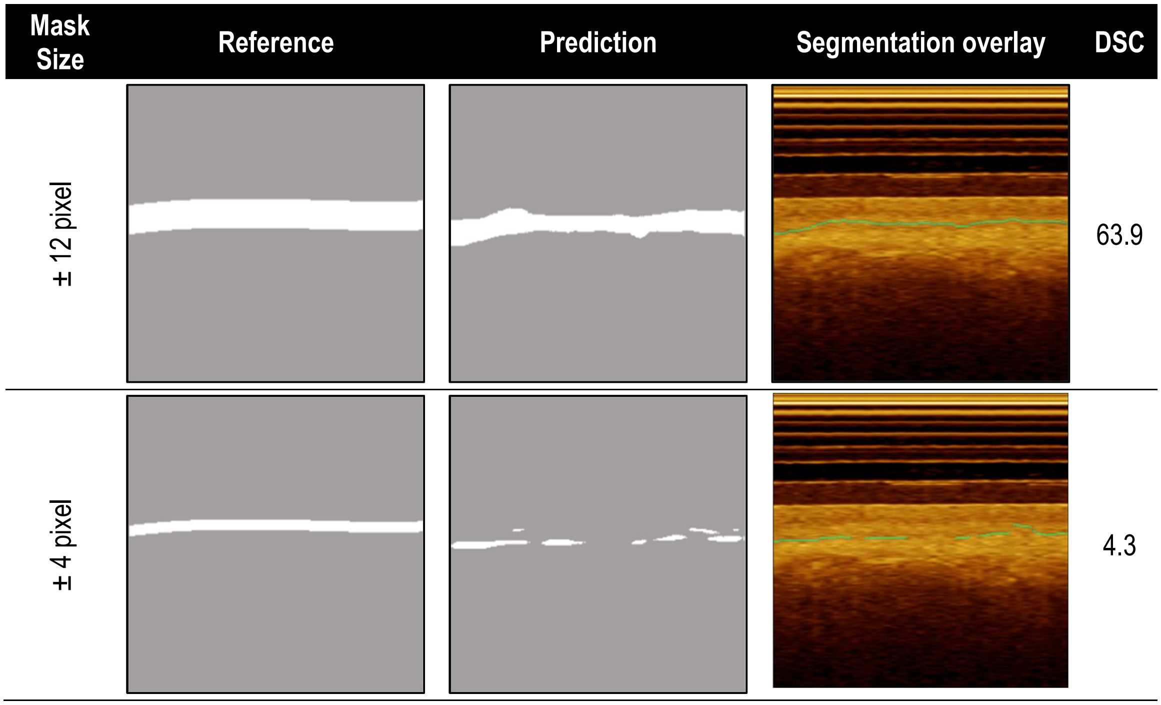


Figure 4.26. Evaluation of basement membrane predictions; Scenario 1: increased DSC after fine-tuning

Scenario 2: Decrease in DSC after fine-tuning

Examining the results of the tile exhibiting the biggest loss in DSC after fine-tuning allows insight when continued training results in worse predictive power. In this tile, the pre-trained model calculated a DSC 63.9, but dropped to 4.3 after fine-tuning. Shown in Figure 4.27, fine-tuning with ± 4 pixel masks results in the undesirable loss of a continuous segmentation, which is present in the pre-trained model. While some of the discontinuity is handled by the post processing algorithm, and the DSC is much improved over the raw prediction, this tile demonstrates a scenario which is not well handled by current methods.



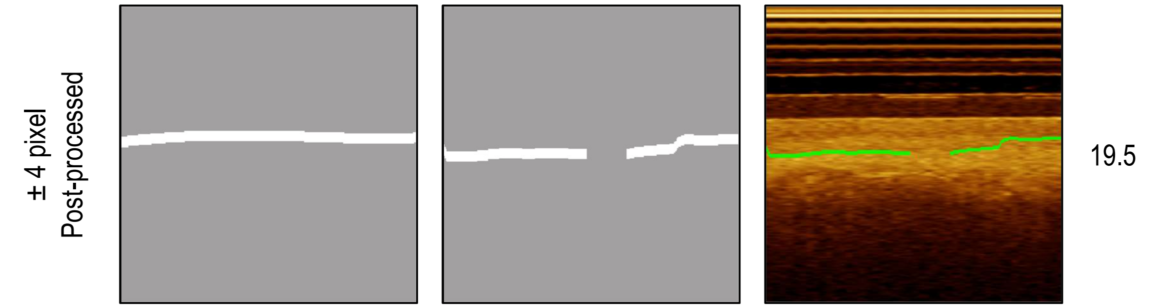


Figure 4.27. Evaluation of basement membrane predictions; scenario 2: Decrease in DSC after fine-tuning

Further analysis of the tile yields another difference between the reference and prediction, contributing a larger influence to the lower DSC. Beyond a break in the prediction, the boundary is also shifted along the a-line axis, as the network has identified the basement membrane to be deeper than the reference. Analysis of the input image without any overlaying annotations (Figure 4.28), reveals two structures (identified by white arrows). In the reference, the superior structure has been identified as the basement membrane, while the inferior structure was identified the network. In this case, insufficient contextual information is available within the tile to make an assessment, but evaluation of neighbouring tiles allows for identification of the superior structure as correct.

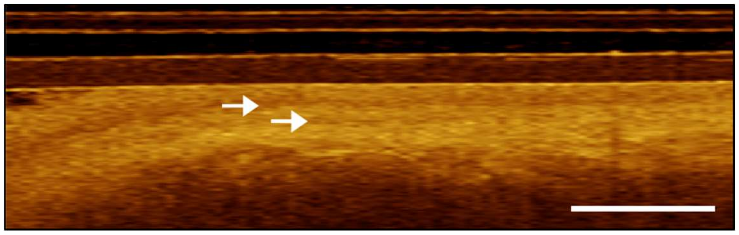


Figure 4.28. Layered tissue structures (identified by white arrows) confounding basement membrane segmentation network predictions. Scale bar 1mm.

Demonstrating Network Success

Results for rater-prediction, rater-reference and inter-rater calculations of the DSC (Table 4.12), PA (Table 4.13), NA (Table 4.14) and pixel error mean and SD (Table 4.15) are presented below. The network prediction metrics are calculated using network model C. Raters that did not segment common slices are marked by N/A.

Table 4.12. Rater-prediction, rater-reference and inter-rater calculations of the DSC for basement membrane segmentation network. N/A indicates raters that did not intersect annotations.

DSC	Reference	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Prediction	82.9	75.9	65.5	78.9	76.7	76.7
Reference		77.7	63.8	74.7	74.7	73.0
Rater 1			57.3	64.2	78.1	67.8
Rater 2				69.3	65.0	N/A
Rater 3					77.1	75.6
Rater 4						N/A

Table 4.13. Rater-prediction, rater-reference and inter-rater calculations of PA for basement membrane segmentation network. N/A indicates raters that did not intersect annotations.

PA (%)						
Prediction	0.96	0.95	0.90	0.96	0.98	0.89
Reference		0.97	0.89	0.95	0.95	0.88
Rater 1			0.91	0.93	0.96	0.87
Rater 2				0.96	0.89	N/A
Rater 3					0.98	0.92
Rater 4						N/A

Table 4.14. Rater-prediction, rater-reference and inter-rater calculations of NA for basement membrane segmentation network. N/A indicates raters that did not intersect annotations.

NA (%)	Reference	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Prediction	0.82	0.71	0.70	0.73	0.85	0.65
Reference		0.85	0.69	0.68	0.74	0.57
Rater 1			0.67	0.64	0.06	0.50
Rater 2				0.76	0.66	N/A
Rater 3					0.38	0.76
Rater 4						N/A

Table 4.15. **Rater-prediction, rater-reference and inter-rater calculations of the mean and SD of pixel error for basement membrane segmentation network. N/A indicates raters that did not intersect annotations.**

μ ± σ	Reference	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5
Prediction	1.15 ± 1.55	1.53 ± 2.06	1.69 ± 2.84	1.36 ± 1.73	1.30 ± 1.59	0.74 ± 0.94
Reference		1.55 ± 1.86	1.80 ± 2.75	1.65 ± 1.89	1.28 ± 1.35	0.93 ± 1.14
Rater 1			2.36 ± 3.72	2.16 ± 2.92	1.57 ± 1.75	1.21 ± 1.52
Rater 2				2.34 ± 3.44	1.41 ± 2.09	N/A
Rater 3					1.87 ± 1.79	0.93 ± 1.26
Rater 4						N/A

As with the epithelial surface segmentation task, the network reference-prediction metrics (DSC = 82.8, PA = 96% NA = 82% μ = 1.15, σ = 1.55) lie within or above the distribution of the inter-rater metrics, demonstrating that the inaccuracies of the network predictions are no worse than inter-rater disagreements.

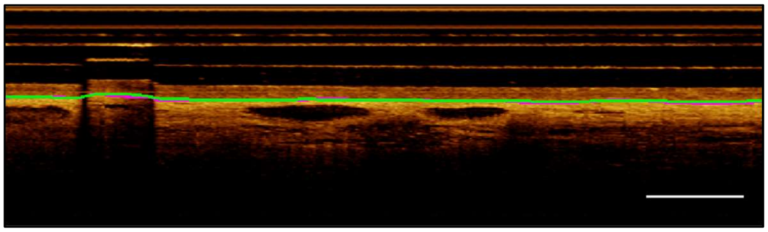
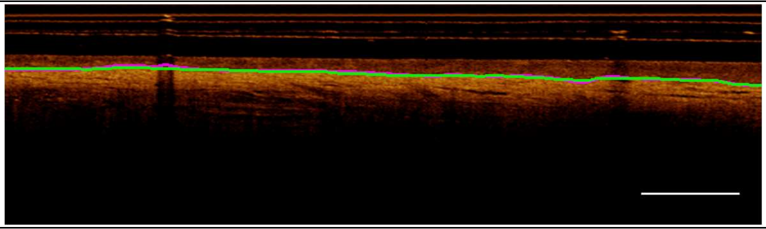
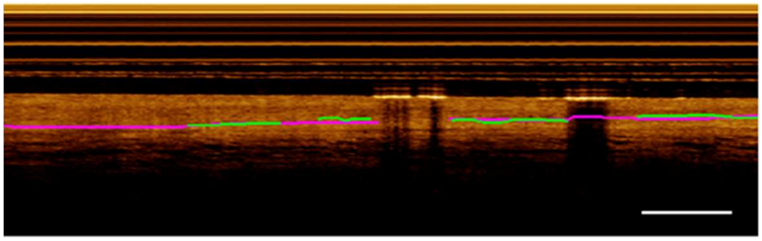
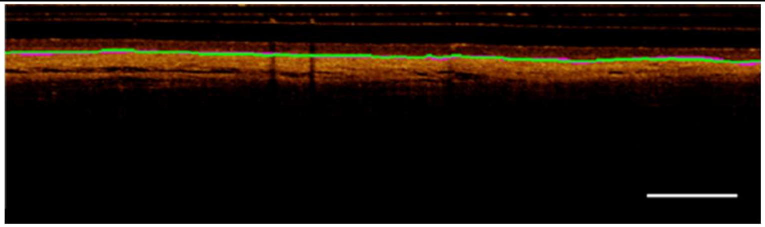
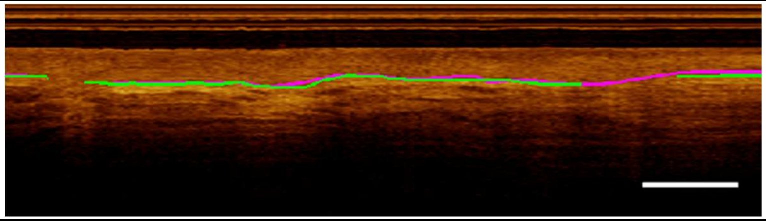
Challenges & Limitations

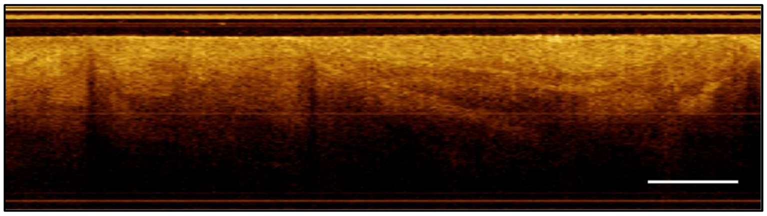
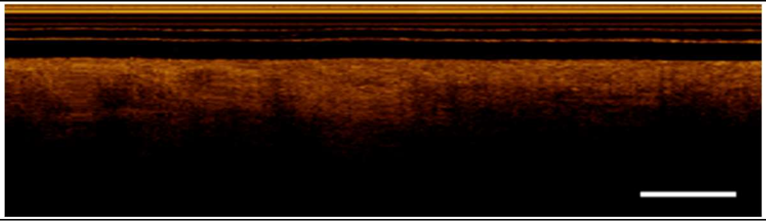
The cellular disorganization characteristic of cancerous and pre-cancerous oral lesions is difficult to quantify using manual annotations, and certain cases required the raters to use contextual information present in the entire longitudinal slice. However, this information is lost during the image tiling step, limiting the predictive power of the networks.

A goal of this task was ensuring generalization across the variety of tissue types present in the oral mucosa. The network was designed to be robust with respect to the variety of tissue types arising from different imaging sites, as well as pathologies. However, as discussed in Section 3.1, the dominant imaging site is the tongue (which aligns with the most common oral cancer presentation [70]), and the dominant pathologies are mild dysplasia, followed by OSCC. For this task, cohorts were not partitioned by site or specific pathology but were instead distributed based on the presence of the basement membrane, as identified by the reference. Accordingly, only a subset of the pathologies and imaging sites were represented in the test set, limiting the ability to provide a comprehensive analysis of network generalizability; a complete summary of pathologies and sites comprising the test set is included in Appendix C.

Example segmentations for available sites and pathologies are detailed in Table 4.16, including two contralateral sites and three pathologic sites.

Table 4.16. Sample segmentations of available pathologies and imaging sites. DSC calculated after post-processing. CIS – carcinoma in situ, OSCC – oral squamous cell carcinoma.

Imaging Site (Pathology)	Metrics DSC PA (%) NA (%) μ ± σ	Segmentation (scale bars 1mm; reference, prediction)
Lateral Tongue (Contralateral)	91.1 99.9 0.0* 0.79 ± 0.85	
Ventral Tongue (Contralateral)	90.7 100.0 N/A** 0.84 ± 0.75	
Buccal Mucosa (Grade 1 Dysplasia)	84.3 95.0 47.0 1.05 ± 1.11	
Ventral Tongue (Grade 2 Dysplasia)	88.9 99.7 0.0* 0.98 ± 0.89	
Lateral Tongue (Grade 3 Dysplasia)	78.5 97.3 66.2 1.66 ± 1.64	

Gingiva (CIS)	100.0 N/A† 100.0 N/A†	
Ventral Tongue (OSCC)	100.0 N/A† 100.0 N/A†	

* NA = 0 when no TN samples (prediction has no negatives)

** NA = N/A cannot be calculated when no negative samples exist in reference or prediction

† PA and pixel error metrics cannot be calculated when no reference or prediction exists.

As expected, the contralateral samples present high reference-prediction agreement, as a continuous membrane can be visualised. The pathologic cases contain a wider variation of metrics, with total agreement in the CIS case, where no basement membrane can be identified.

The segmentation of the grade 3 lateral tongue presents the lowest DSC score. This may be a result of poor representation of lateral tongue presenting grade 3 dysplasia (n = 2). Both samples were in the test cohort. Moreover, only seven examples of grade 3 dysplasia where present in the entire dataset, and six of these cases were in the test cohort. As it is, any similarity provides evidence that the network exhibits rudimentary generalizability

Additionally, noisy labels may contribute to low metrics; noisy labels occur due to rater error, for example missed identification of the basement membrane. Re-evaluation of the unannotated slice of a ventral tongue sample shown in Figure 4.29, panel a, could reasonably introduce annotations similar to the prediction, shown in panel b. These types of errors motivate the development automated methods, reducing the likelihood of human error.

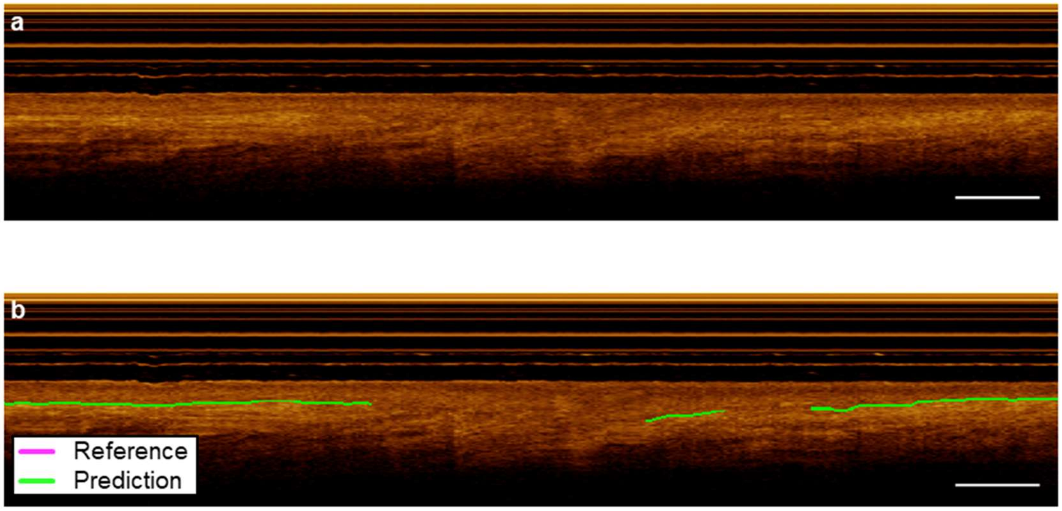


Figure 4.29. (a) Example of empty basement membrane reference contradicted by (b) network predicted boundary.

A persistently low NA compared to PA indicates that the model under predicts regions of basement membrane (i.e. misses regions identified in the reference set). In the context of oral cancer, it may be more favourable to over-predict suspicious regions. However, this remains an avenue for continued development through identification of regions more susceptible to missing predictions. For example, the buccal mucosa sample presenting with grade 1 dysplasia features the lowest NA, indicating under-prediction of the basement membrane. Poor representation of pathologies is pervasive this dataset and may account for the low metrics (n=3 samples of buccal mucosa presenting with grade 1 dysplasia). However, evaluation of the unannotated frame (Figure 4.30, panel a) reveals that this sample presents regions of lower intensity, coinciding with regions of missed annotations (panel b), highlighted by white arrows). Additionally, imaging artifacts (identified by white stars) also demonstrate regions of missing predictions. Inconsistent reference methods can also be observed through comparison the two artifacts, where a prediction has been added for the far-right artifact, but no reference exists for the left artifact; this is another example of noisy labels.

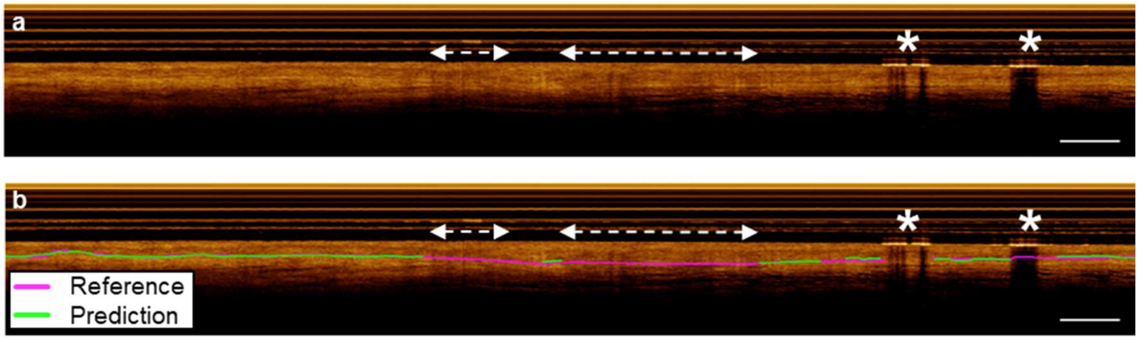


Figure 4.30. (a) Unannotated sample; (b) corresponding reference and incorrect predictions. White arrows indicating regions of lower intensity, providing possible explanation for disagreement. White stars highlight imaging artifacts which may also contribute to poor predictions.

4.5. Detection of Imaging Artifacts

This section describes the fourth DL network, designed to detect imaging artifacts arising from air bubbles in the imaging probe's protective layers, as well as any markers that have been printed on the sheath. Artifacts cause a decrease or complete loss of image intensity and may confound other steps of the pipeline. While predictions made with this model are not used as inputs to other networks, projecting the results into the *en face* view allows for quick comparison and may provide insight about regions where loss of basement membrane is due to imaging artifacts rather than dysplastic changes.

4.5.1. Methods

Dataset Preparation

Training, tuning and testing cohorts were generated by dividing longitudinal OCT slices into 128x128 pixel tiles, with 64 pixel overlap. Using the bubble and sheath marker reference annotations, tiles were classified according to the decision tree in Figure 4.31. Tiles with small artifacts (<30% of the tile) occurring at the edge of the tile (borderline tiles) were discarded, and remaining tiles were assigned to class 0 (containing no artifact) or class 1 (containing an artifact).

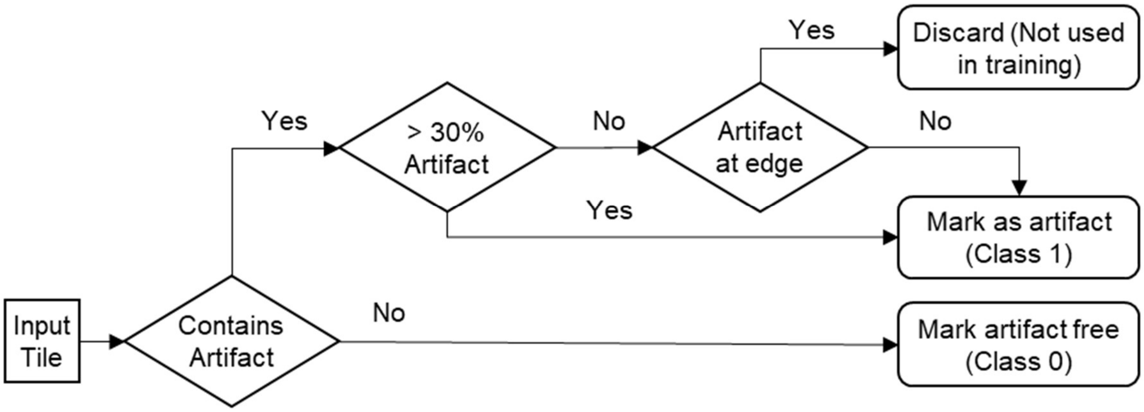

Input Tile
Contains Artifact
Yes
No
> 30% Artifact
Yes
No
Artifact at edge
Yes
No
Discard (Not used in training)
Mark as artifact (Class 1)
Mark artifact free (Class 0)

Figure 4.31. Image artifact tile class assignment and exclusion decision tree.

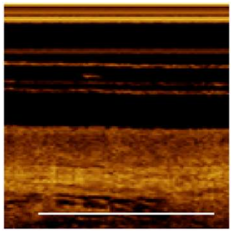
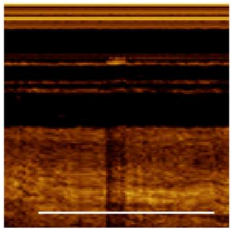
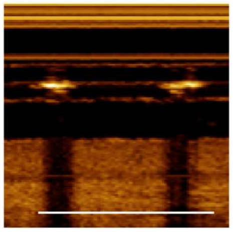
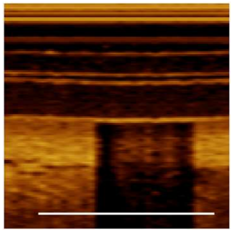
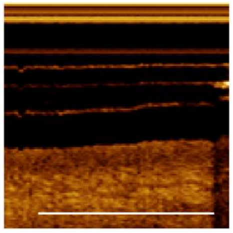
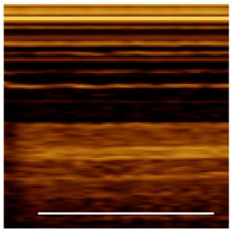
Exclusion of borderline tiles was permissible as tile overlap ensured artifact presence in neighbouring tiles. The distribution of tiles per cohort is summarized in Table 4.17, and representative examples of tiles from class 0, class 1, and exclusions are shown in

Table 4.18. Z-score normalization was applied to all cohorts.

Table 4.17. Distribution of tiles used for development of image artifact detection network.

Cohort	No. Tiles (% of cohort)			No. Patients
	Class 0	Class 1	Excluded	
Train	16,059 (93.6)	1,095 (6.4)	402 (0)	39
Tune	2,229 (91.8)	198 (8.2)	84 (0)	9
Test	2,179 (92.9)	166 (7.1)	58 (0)	11
Total	20,467	1,459	544	

Table 4.18. Representative samples of tiles used for development of image artifact detection network.

Class	Representative Tile(s). Scale bars 1mm
Class 0	
Class 1	  
Excluded	 

Network Definition & Parameters

Similar to the FOV task, a custom CNN was used for artifact detection. However, the more complex presentation of the target features informed that a deeper network was necessary to generate accurate predictions. Network topology is shown in Figure 4.32. Further details on network layers can be found in Appendix B, Table B4. Hyperparameters implemented for network training are summarized in Table 4.1.

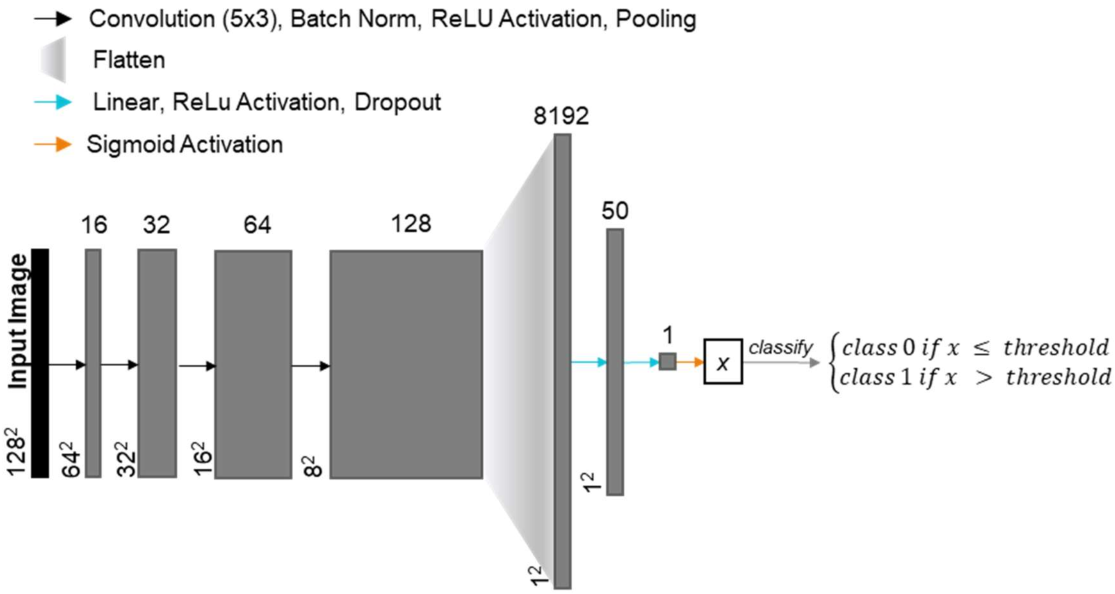
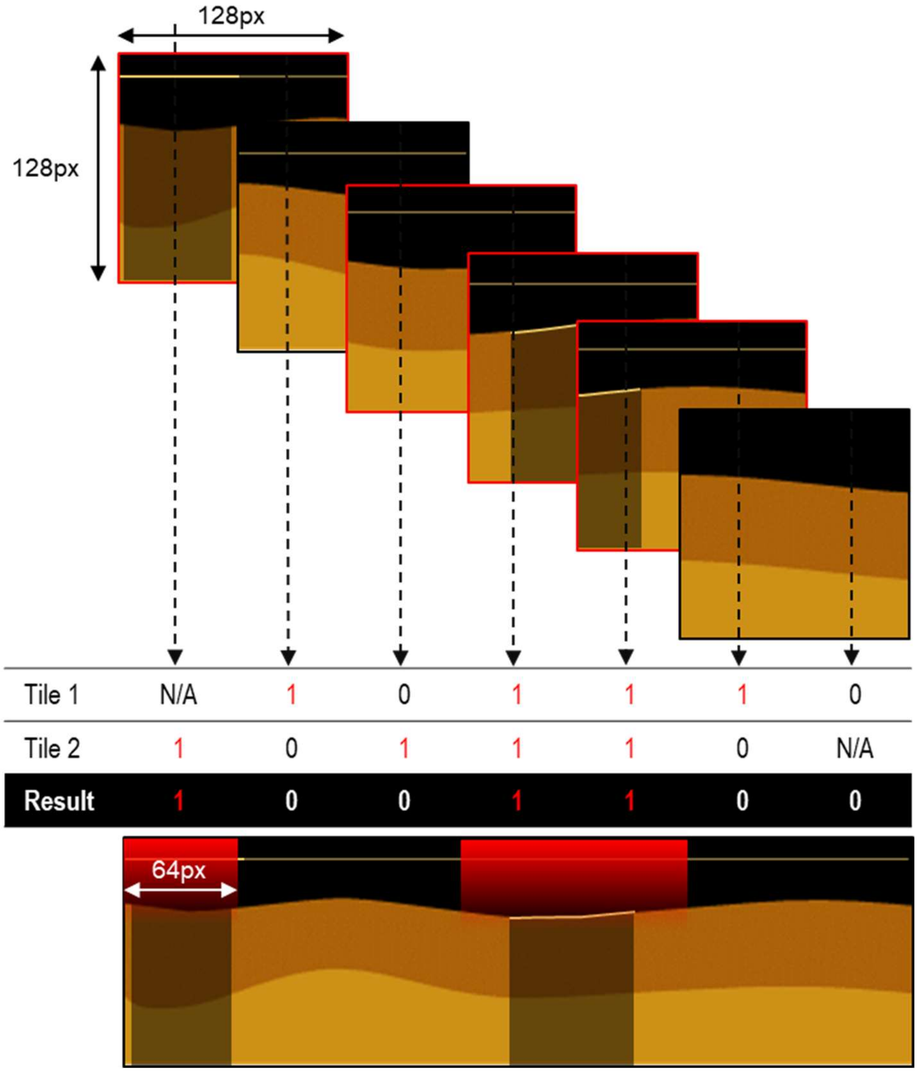

Convolution (5x3), Batch Norm, ReLU Activation, Pooling
Flatten
Linear, ReLu Activation, Dropout
Sigmoid Activation
Input Image
128^2
16
64^2
32
32^2
64
16^2
128
8^2
8192
1^2
50
1^2
1
x
classify
$\begin{cases} class\,0\; if\; x \le threshold \\ class\,1\; if\; x > threshold \end{cases}$

Figure 4.32. Network Topology for artifact detection network. Arrows represent layer operations (defined, top right), grey boxes represent feature maps, labelled with number of feature maps (top) and image dimension (left).

Post-Processing

Tile predictions were stitched into their longitudinal frames. To increase the resolution beyond the tile wise classification of 128 pixels wide, overlapping tile regions were assessed to determine the presence of an artifact, allowing for a resolution of 64 pixels. As shown in Figure 4.33, neighboring tiles must agree on artifact presence in order to propagate into the final frame. Edge tiles were exempt from this requirement. In the schematic, tiles predicted as Class 1 are identified by a red border, and Class 0 predictions have black borders.



Tile 1	N/A	1	0	1	1	1	0
Tile 2	1	0	1	1	1	0	N/A
Result	**1**	**0**	**0**	**1**	**1**	**0**	**0**

Figure 4.33. Schematic of post-processing for artifact detection network. Network prediction tiles are shown at top; Class 1 predictions are identified by red borders, Class 0 predictions are identified by black borders. Projection of overlapping tiles must agree for artifact to be identified in longitudinal slice (shown at bottom).

4.5.2. Results

Classification metrics for the artifact detection network are reported in Table 4.19. Tile metrics are calculated after applying a binary threshold to the output of the sigmoid layer, which is calculated using the Youden J index of the tuning set. ROC and PRC curves for the test set are presented in Figure 4.34. While these curves represent less powerful predictors than the FOV detection model, they represent the significant improvement over an unskilled network.

Table 4.19. Classification metrics for artifact detection network

Metric	Value
Threshold	0.38
Balanced Accuracy	75.7%
Sensitivity	99.1%
Specificity	52.3%
AUC	93.9
mAP	68.1

		PREDICTED CLASS	
		CLASS 0	CLASS 1
ACTUAL CLASS	CLASS 0	2109	63
	CLASS 1	18	69

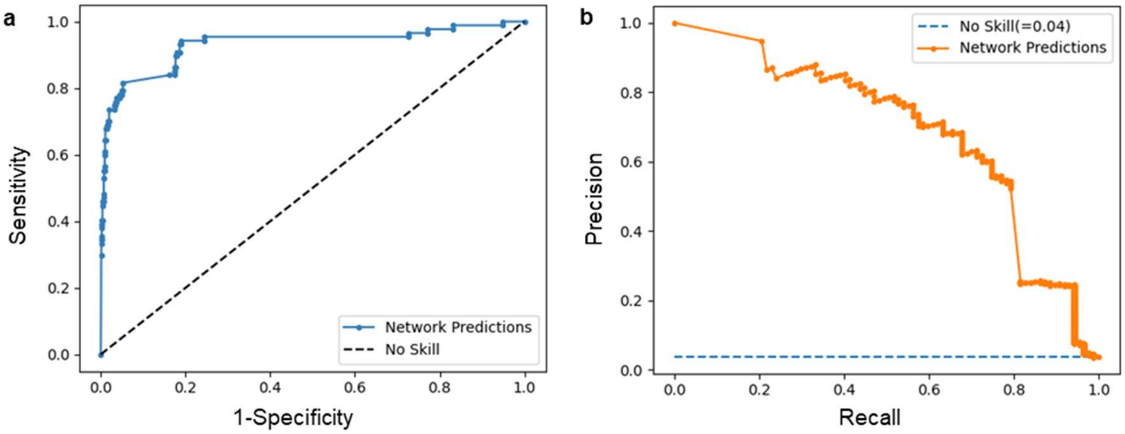


Figure 4.34. Artifact detection tile classification (a) ROC curve; (b) PRC curve. Baseline PRC = 0.04.

4.5.3. Discussion

The artifact detection CNN was moderately successful in identifying artifacts correctly. Examination of incorrectly predicted tiles elucidated several confounding features contributing to misclassifications. Examples of tiles incorrectly identified to be artifacts (Type 1 error, n = 63) are shown in Figure 4.35.

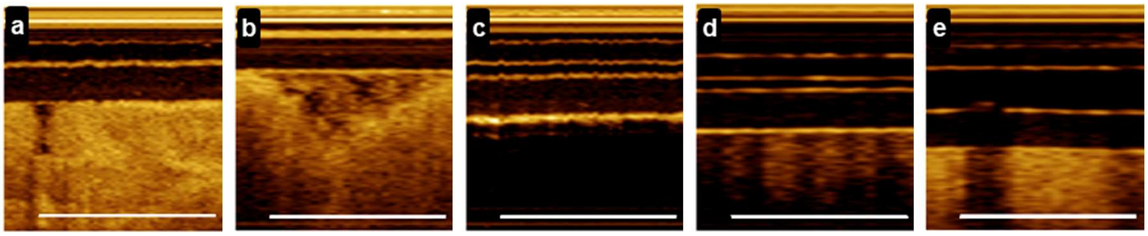


Figure 4.35. Type 1 errors in artifact detection network predictions. Scale bars 1mm.

Tiles containing the tissue holes (panel a), tiles with mucous (panel b), tiles with no tissue data (panel c) were commonly misclassified. While cases such as panel a and b are uncommon, cases like panel c are more prevalent and are commonly a result of poor probe to tissue contact. This occurs most often at the edge of a pull back and particularly in difficult to reach sites of the oral cavity. However, these types of misclassification may actually serve as a benefit, as they present abnormalities that may contribute to misleading or mistaken labels in the previous networks. Without image context, poor OCT quality is the apparent contributor to the misclassification of panel d. While this is an undesirable outcome, evaluation of the entire slice (Figure 4.36; Figure 4.35, panel d outlined) reveals a nearby artifact that may have resulted in the decreased intensity. The classification error of Figure 4.35, panel e is attributed to an incorrect label during the annotation process.

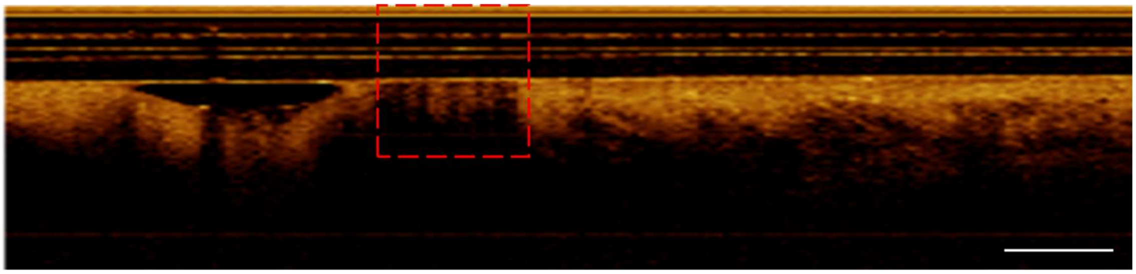


Figure 4.36. Exploring incorrect tile exclusion from artifact detection model; incorrect tile overlay (red box) on section of whole slice. Scale bars 1mm.

Examples of incorrectly excluded tiles (type 2 error, n = 18) are shown in Figure 4.37.

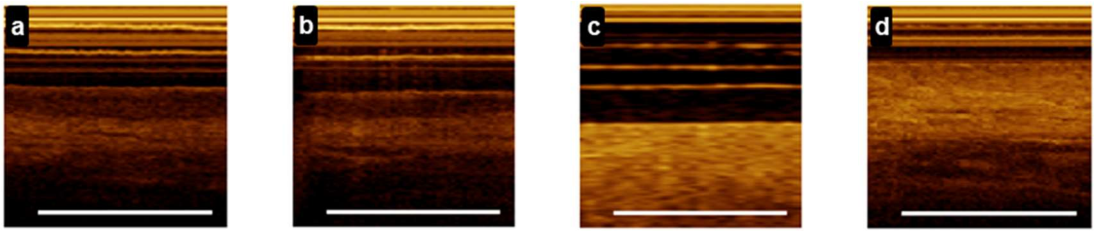


Figure 4.37. Type 2 errors (false negative) in artifact detection network predictions. Scale bars 1mm.

Examination of the FN predictions revealed that 13 tiles were from the same pullback (Figure 4.38, artifact annotated). It is clear when visualizing a broader section that a large artifact encompasses entire tiles (Figure 4.37, panel a and b). The vastness of the air bubble, compounded with the proximity of the sheath and window tube (restricting the height of the bubble, and thus the path length of the laser) result in moderate quality OCT data, and minimal shifts along the a-line axis. As such, this data was comparable to artifact free data. Additionally, tiles that encompassed the transition into and out of the artifact were correctly identified. Panels c and d in Figure 4.37 were identified as misclassified during rater annotation.

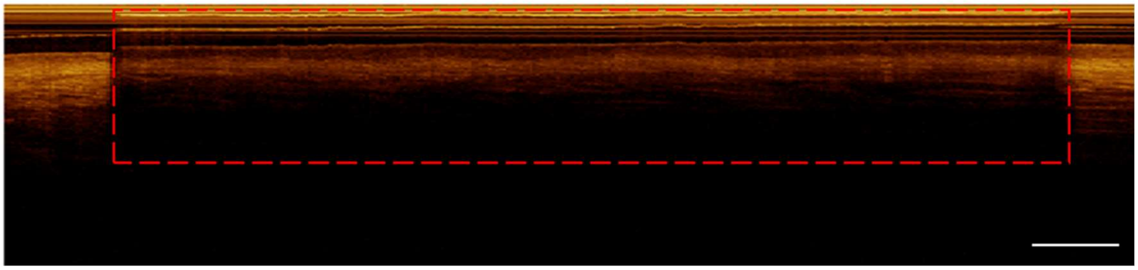


Figure 4.38 Exploring incorrect tile inclusion of artifact detection model; incorrectly identified region (red box) section of whole slice. Scale bars 1mm.

The smaller size of training tiles was intentionally selected as the imaging artifacts targeted by this task are often the most identifiable by regions of increased specular reflection in the sheath. Additionally, post-processing could be applied by averaging neighbouring tile overlaps such that artifacts could be located with 64-pixel precision, as tile-wise classification generated a more granular result. However, as observed, the shallow depth resulted in occasional misclassification of samples with no tissue contact. Experimentation was performed on 256x256 pixel tiles to allow for more context during training. However, worse outcomes were observed, resulting in misclassifications of tiles with loss of image stratification, hypothesized to be a result of image shadows more prevalent in pathologic samples.

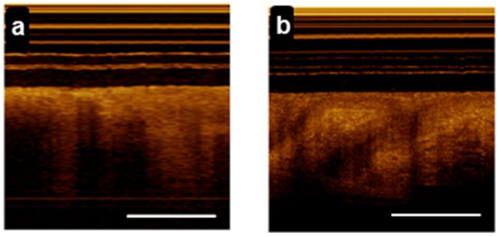


Figure 4.39. Type 1 errors for artifact detection using 256x256 pixel tiles.

4.6. Whole Volume Analysis

This section presents the development of the epithelial thickness maps. Synthesis of the DL predictions was necessary to allow for rapid and intuitive localization of potentially pathologic regions.

Two pullbacks were excluded from the development of the preceding networks to serve as an external test cohort. A true external cohort is not possible in this retrospective study and other research groups have not generated imaging in a way that is comparable without requiring substantial domain transfer. In lieu, a single patient with confirmed dysplasia (grade 2) on the ventral tongue was isolated; data from this patient excluded from the network discovery set.

4.6.1. Methods

One pullback was taken across the lesion and a second across the contralateral site. 504 longitudinal slices were generated from each pullback; this number is determined by rotational speed of the probe, as well as the number of captured a-lines per longitudinal slice. Each slice was passed through the DL networks pipeline sequentially. First, slices outside the imaging FOV were discarded (per Section 4.2). Accepted slices were then analyzed to first localize the epithelial surface (per Section 4.3), then detect and localize the basement membrane (per Section 4.4), and finally identify imaging artifacts that may contribute to misinformation in the prediction (per

Section 4.5). Longitudinal slice predictions were combined to reconstruct a 3D volume detailing the epithelial layer and imaging artifact. Epithelial thickness maps are corrected for refractive index of water (n = 1.33), which closely matches that of tissue.

4.6.2. Results

Shown for the contralateral volume (Figure 4.40) and the dysplastic volume (Figure 4.41), synthesis of the 3D prediction data volumes was used to generate an epithelial thickness map (panel b), a broken or missing basement membrane map (panel c) and an artifact identification map (panel d), each overlaid onto respective *en face* projections. Panel a shows an unmarked *en face* projection for reference. In MATLAB, a small Gaussian blur has been applied to the epithelial thickness maps (kernel size = [10, 10]), and artifact identification map (kernel size = [5, 15]) to smooth edges. Expanded view of these figures is included in Appendix D.

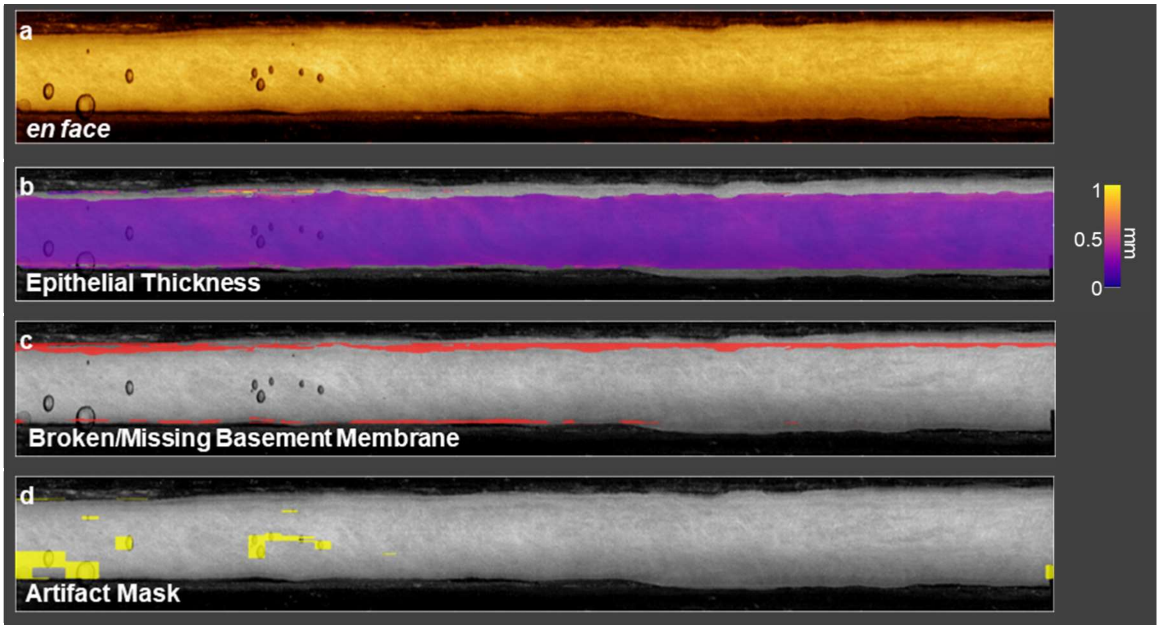


Figure 4.40. En face of OCT pullback of (presumed normal) contralateral lateral tongue (a) unannotated; (b) with superimposed epithelial thickness map; (c) with superimposed artifact locale mask.

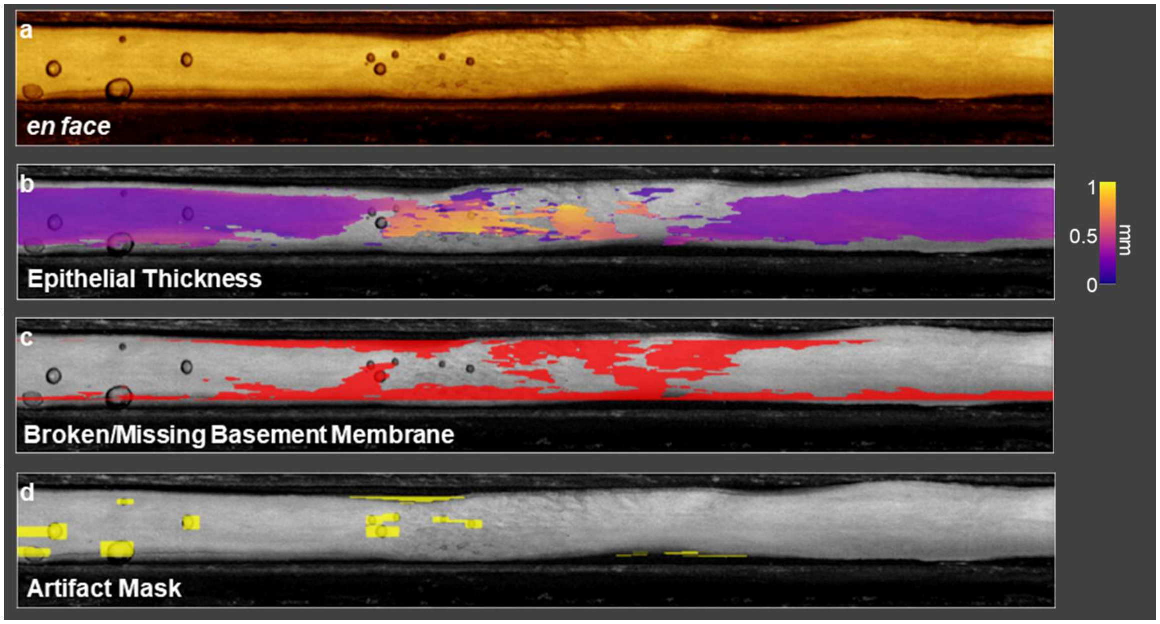


Figure 4.41. En face of OCT pullback of lateral tongue with grade 2 dysplasia (a) unannotated; (b) with superimposed epithelial thickness map; (c) with superimposed regions of missing or broken basement membrane; (d) with superimposed artifact locale mask.

4.6.3. Discussion

The differences between the two epithelial thickness maps demonstrate ability to discriminate presumed normal tissue from moderate dysplasia (grade 2). This highlights the opportunity for automated depth analysis of OCT data, wherein differences in tissue content are non-obvious in the *en face* view, and the quantity of longitudinal slices renders manual assessment unmanageable. Additionally, presenting depth information in the *en face* view allows for qualitative image interpretation.

The epithelial thickness map of the contralateral volume (Figure 4.40, panel b) displays a near uniform epithelial thickness, aligning with the expected biology of the tissue. The DL pipeline was able to accurately detect the imaging FOV, and generate precise segmentation of the epithelial surface and basement membrane. This observation is further supported by isolating a single longitudinal scan (Figure 4.42, panel a), which demonstrates accurate epithelial layer identification. A slight thickening of the epithelium can be observed at the top left corner of the thickness map. However, analysis of the longitudinal slice in this area demonstrates acceptable segmentation, as the tissue structures are not captured by the OCT system at FOV fringes (Figure 4.42, panel b). A stricter FOV classifier may better reject these slices.

Figure 4.40, panel c details the predictions of the artifact detection model. While granular, artifacts have been correctly identified, including a partial sheath marker in the lower right corner. The application of tile-wise classification of artifact presence results in aliasing in identification of artifacts, with a resolution of 64 pixels along the pullback direction, and one pixel along the azimuthal direction. This results in some off apparent off-target predictions, where a small section of the artifact may be present at the edge of the tile. Future steps could involve performing artifact detection in the *en face* projection, allowing for more precise localization.

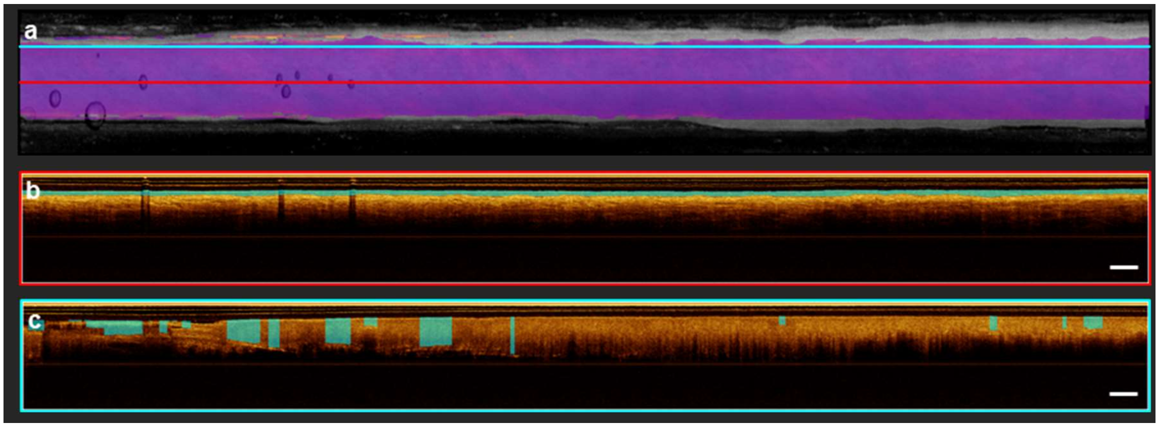


Figure 4.42. Annotated longitudinal slices of contralateral pullback from (a) center FOV imaging; (b) fringe FOV. Scale bars 1mm.

Analysis of the dysplastic epithelial thickness map (Figure 4.41, panel b) shows the thickening and loss of the basement membrane characteristic to oral dysplasia. Approaching the lesion from the left, the map indicates a slow thickening of the epithelial layer. In comparison, approaching the lesion from the right indicates a sharp transition from stratified into non-stratified tissue. Evaluation of a longitudinal slice pulled from the center of the imaging FOV supports this observation (Figure 4.43). While some inconsistencies and marked changes in the depth of the identified basement membrane are present in the segmentation boundary, the general trend of abnormal tissue towards the center of the tissue, versus the more normal appearing tissue towards the image edges is captured. Moreover, comparison of the contralateral versus dysplastic epithelial layer masks allows for observation of slightly thickened epithelium even in well stratified tissues, supporting the expected biological differences between cancerous and potentially cancerous oral tissue.

Evaluation of the image artifact mask (Figure 4.41, panel c) supports conclusions drawn in the contralateral case. It should be noted that the relatively small size of the artifacts in these samples has not caused observable changes to the epithelial thickness maps, but analysis with more artifact presence may demonstrate otherwise.

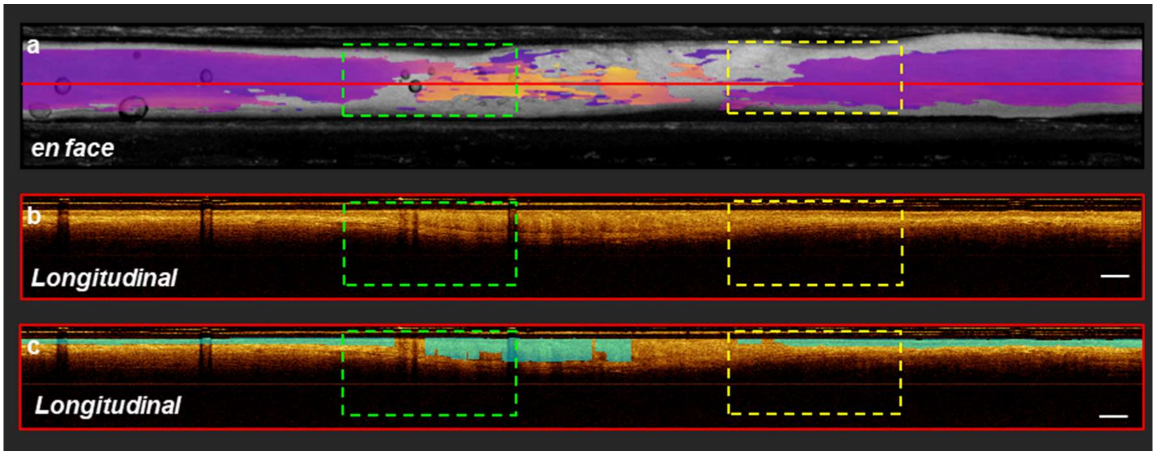


Figure 4.43. Annotated longitudinal slice of dysplasia grade 2 pullback over center FOV. Scale bar 1mm.

4.7. Discussion

Identification of the most pathologic area is a crucial step in early oral cancer detection. Localization of the most pathologic area during biopsy procedures may reduce the amount of false negatives; additionally, accurate assessment of lesion margins can assist in reducing the amount of excess tissue discarded during surgical resection, a necessary step to prevent recurrences. OCT has demonstrated clinical utility in identifying these regions of interest, but interpretation of OCT is a challenging task, both due to the vast amount of data acquired, but also the expertise required to interpret the data.

In this chapter, a pipeline for automated segmentation of oral OCT is presented. Pre and post processing steps are developed in MATLAB, bookending four DL networks developed in Python 3.6. Pre-processing methods included automated contrast and brightness balancing of longitudinal slices, and dividing slices into small tiles for network development. DL Networks were trained to be independent and order invariant, except for the FOV detection network, which must be completed first. With current organization, the second and third steps of the pipeline provide segmentations of the surface and basement membrane, respectively. Segmentation predictions are generated with customized u-nets, adjusted to suit the intricacies of applying area segmentation networks for boundary detection. The last step of the pipeline identifies regions containing imaging artifacts. Post-processing methods were implemented to clean and smoothen network segmentation predictions, as per-pixel classification methods often result in spurious predictions and jagged edges. Finally, epithelial layer maps were generated through combination of epithelial surface and basement membrane segmentations, providing information about changes to epithelial thickening and stratification. Depending on the amount of longitudinal slices within the FOV, processing time for an entire OCT volume is about 8 minutes, comprising 100-130 seconds for DL processing, and 300 seconds for the post-processing steps.

Applying DL methods towards analysis of oral OCT features has demonstrated successful results; errors between segmentation and reference information were no greater than inconsistencies and disagreements between expert raters. Training segmentation networks to run independently prevented propagating errors through the analysis pipeline, while allowing easier implementation of domain-transfer tasks, such as implementation in lung and cervical OCT datasets.

4.7.1. Difficulties in OCT interpretation

As discussed in Chapter 3, annotations were generated by 6 raters: five experienced OCT researchers, and one undergraduate trainee. While all raters underwent the same training and standardization sessions, inter-rater evaluation revealed that the lack of OCT familiarity of the undergraduate trainee often resulted in incorrect annotation labels, particularly in the more complicated basement membrane segmentation task. Consequentially, this rater was excluded from generation of the consensus data set and evaluation of network success. While a small sample size is not sufficient to draw meaningful conclusions, this inconsistency further supports the need for automated systems to analyze OCT data, as clinicians may not be familiar with interpreting images generated using OCT.

4.7.2. Network Limitations

While an ideal segmentation network would be generalizable to tissue across the oral mucosa, the networks presented in this research are unavoidably biased towards the more prevalent presentation of oral cancer of the tongue. While the presentation of the epithelial surface remains consistent, the results in Section 4.4.3 demonstrate that changes in imaging site and pathology may influence network predictions of the basement membrane. However, at this time, there are insufficient examples to demonstrates statistical significance.

Moreover, uncommon tissue features (e.g. Figure 4.44) may still confound the network. In this tissue (floor of mouth), a thick, highly scattering region can be observed on the middle-left region of the longitudinal slice. Comparing the network prediction (panel a, shown as the raw prediction for easier visualization) to the reference annotation (panel b) reveals undesired gaps in the segmentations which is not representative of the tissue organization. Such features, believed to be a keratin structure on the tissue surface, are not common to the training set. However, it is probable that the network could learn to generate correct predictions, if exposed to more examples of similar data. Such limitations are common in DL tasks, as algorithms excel at tasks with repetitive, consistent data, and struggle with unknown information or differing presentations.

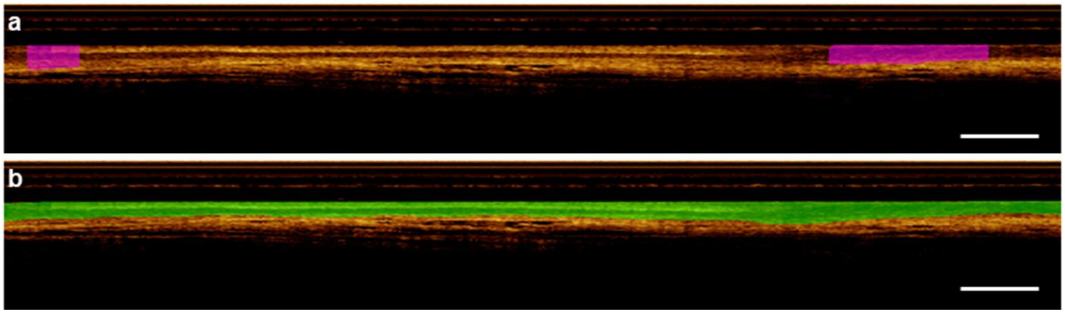


Figure 4.44. Example of poor segmentation prediction resulting from uncommon image features. (a) Raw network prediction mask; (b) reference annotation mask. Scale bars 1mm.

Limitations that prevent clinical deployment of this research include: (1) an absence of cross validation methods, (2) no external cohort evaluation. The lack of cross-validation is primarily attributed to the manual input required to generate the network evaluation cohorts, which is a consequence of the unequal number of longitudinal slices selected per patient, compounded with uneven patient prevalence. As discussed, the novelty of this research limits the number of similar data sets and consequently the ability to source external cohorts. To allow for clinical deployment, future work must include implementation of cross validation methods, reporting of network confidence intervals and confirmation of network generalizability on an external dataset.

Chapter 5. Future Work & Conclusion

5.1. Clinical Workflow Adoption

This work presents a promising comprehensive tool which may have utility as an adjunct for biopsy guidance and surgical margin identification. However, the identified limitations of no cross-validation methods, nor evaluation of an external cohort are significant barriers preventing clinical adoption. Cross-validation methods are fundamental to generate confidence intervals and instill trustworthiness in the model. Conversely, an external cohort necessary to confirm reproducibility and generalizability.

Additional barriers to clinical utility include dependencies on expensive hardware and a large working memory to process data at a reasonable rate. These limitations may be mitigated with efficient programming and translation into a faster programming language, such as C++, may yield further improvements in execution time.

5.2. Correlation of OCT Volumes with Biopsy Data

Epithelial thickness is an established biomarker for histopathologic assessment of benign, potentially cancerous, and cancerous conditions of the oral cavity [43]. Creation of epithelial thickness maps from OCT data may reveal similar information to histology, but annotations of entire samples are necessary to draw meaningful conclusions. The application of this thesis may serve as the foundation to rapidly generate such annotations, which in turn could be used to generate relationships between the tissue features and the pathology of the sample, confirmed by biopsy.

5.3. Development of DL Methods for Other OCT Data

The helical scanning pattern, combined with the small fiber-optic probe used to collect the data for this work is particularly suited for imaging small luminal organs. The success of the networks presented in this thesis motivate expanding applications into analysis of OCT data from other imaging sites. The BC Cancer OCIL group has amassed a large volume of *in vivo* and *ex vivo* OCT data from pulmonary and gynecological sites. Prior work has suggested that OCT may be used to assess pathologic changes occurring during airway remodelling [71], [72], as well as in the subsurface structures of the cervical canal [73] and fallopian tubes [74]. Analogous challenges to oral OCT motivate limited clinical utility for these applications, with manual rating being time-consuming and suffering from poor repeatability.

Preliminary work is currently underway to explore if (1) the network topologies developed during this thesis can be trained using other clinical data (2) the network weights specific to the oral OCT models can be used for pre-training future networks.

5.4. Conclusion

OCT has the potential to improve the early detection and monitoring of oral cancers, enabling clinicians to intervene at earlier stages when treatment outcomes are generally more favorable. Clinical adoption of this technology requires image analysis tools to provide rapid and reproducible assessment of tissue state during biopsy procedures. However, the large volume of data collected with OCT makes manual annotation intractable.

The overarching goal of this work was to create ancillary software to complement oral OCT and empower clinicians with another modality to make treatment plan recommendations. With this thesis, a novel DL pipeline is presented as a tool to detect and quantify the presence and depth of the epithelial layer in OCT images. Defining the tissue analysis task to be completed through segmentation allows for the creation of a more generalized approach, not restricted to specific oral sites and diseases, when compared to previous classification tasks. Built with two custom CNNs, for FOV and artifact detection, and two modified u-nets, for epithelial layer isolation, this pipeline provides fast and reproducible results. Comparison of inter-rater agreement to network predictions demonstrate as-good-as or better agreement, and evaluation of whole OCT volumes of exemplifies ability to identify pathological progression. The development of a custom OCT segmentation app was an additional benefit of this thesis, simplifying future segmentation tasks of other clinical OCT data.

The contributions of this thesis could promote the integration of oral OCT into the diagnostic workflow for monitoring of oral health, incisional biopsy guidance and identification of tumour margins during surgical excision.

References

[1] “Oral and Oropharyngeal Cancer: Statistics,” *American Society of Clinical Oncology Cancer.Net*, 2022. https://www.cancer.net/cancer-types/oral-and-oropharyngeal-cancer/statistics (accessed Jan. 26, 2023).

[2] Canadian Cancer Society, “Supportive care for leukemia | Canadian Cancer Society,” 2020. https://cancer.ca/en/cancer-information/cancer-types/oral/supportive-care (accessed May 28, 2023).

[3] P. Wilder-Smith *et al.*, “In vivo optical coherence tomography for the diagnosis of oral malignancy,” *Lasers Surg. Med.*, vol. 35, no. 4, pp. 269–275, 2004, doi: 10.1002/LSM.20098.

[4] J. M. Ridgway *et al.*, “In Vivo Optical Coherence Tomography of the Human Oral Cavity and Oropharynx,” *Arch. Otolaryngol. Neck Surg.*, vol. 132, no. 10, pp. 1074–1081, Oct. 2006, doi: 10.1001/ARCHOTOL.132.10.1074.

[5] E. S. Matheny *et al.*, “Diagnosis of oral precancer with optical coherence tomography,” *Biomed. Opt. Express, Vol. 3, Issue 7, pp. 1632-1646*, vol. 3, no. 7, pp. 1632–1646, Jul. 2012, doi: 10.1364/BOE.3.001632.

[6] A. M. D. Lee, L. Cahill, K. Liu, C. MacAulay, C. Poh, and P. Lane, “Wide-field in vivo oral OCT imaging,” *Biomed. Opt. Express*, vol. 6, no. 7, p. 2664, Jul. 2015, doi: 10.1364/boe.6.002664.

[7] M. Pekala, N. Joshi, T. Y. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, “Deep learning based retinal OCT segmentation,” *Comput. Biol. Med.*, vol. 114, p. 103445, Nov. 2019, doi: 10.1016/J.COMPBIOMED.2019.103445.

[8] J. Kugelman *et al.*, “Automatic choroidal segmentation in OCT images using supervised deep learning methods,” *Sci. Rep.*, 2019, doi: 10.1038/s41598-019-49816-4.

[9] L. Fang *et al.*, “Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search,” *Biomed. Opt. Express, Vol. 8, Issue 5, pp. 2732-2744*, vol. 8, no. 5, pp. 2732–

2744, May 2017, doi: 10.1364/BOE.8.002732.

[10] Z. Yang *et al.*, “Connectivity-based deep learning approach for segmentation of the epithelium in in vivo human esophageal OCT images,” *Biomed. Opt. Express, Vol. 12, Issue 10, pp. 6326-6340*, vol. 12, no. 10, pp. 6326–6340, Oct. 2021, doi: 10.1364/BOE.434775.

[11] Y. Gharaibeh *et al.*, “Coronary calcification segmentation in intravascular OCT images using deep learning: application to calcification scoring,” *https://doi.org/10.1117/1.JMI.6.4.045002*, vol. 6, no. 4, p. 045002, Dec. 2019, doi: 10.1117/1.JMI.6.4.045002.

[12] M. Pekala, N. Joshi, T. Y. A. Liu, N. M. Bressler, D. C. DeBuc, and P. Burlina, “Deep learning based retinal OCT segmentation,” *Comput. Biol. Med.*, vol. 114, p. 103445, Nov. 2019, doi: 10.1016/J.COMPBIOMED.2019.103445.

[13] L. Fang *et al.*, “Automatic segmentation of nine retinal layer boundaries in OCT images of non-exudative AMD patients using deep learning and graph search,” *Biomed. Opt. Express, Vol. 8, Issue 5, pp. 2732-2744*, vol. 8, no. 5, pp. 2732–2744, May 2017, doi: 10.1364/BOE.8.002732.

[14] Z. Yang *et al.*, “Connectivity-based deep learning approach for segmentation of the epithelium in in vivo human esophageal OCT images,” *Biomed. Opt. Express, Vol. 12, Issue 10, pp. 6326-6340*, vol. 12, no. 10, pp. 6326–6340, Oct. 2021, doi: 10.1364/BOE.434775.

[15] Y. Gharaibeh *et al.*, “Coronary calcification segmentation in intravascular OCT images using deep learning: application to calcification scoring,” *https://doi.org/10.1117/1.JMI.6.4.045002*, vol. 6, no. 4, p. 045002, Dec. 2019, doi: 10.1117/1.JMI.6.4.045002.

[16] C. S. Lee, D. M. Baughman, and A. Y. Lee, “Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images,” *Kidney Int. Reports*, vol. 1, no. 4, pp. 322–327, 2017, doi: 10.1016/j.oret.2016.12.009.

[17] L. Fang, C. Wang, S. Li, H. Rabbani, X. Chen, and Z. Liu, “Attention to lesion:

Lesion-Aware convolutional neural network for retinal optical coherence tomography image classification,” *IEEE Trans. Med. Imaging*, vol. 38, no. 8, pp. 1959–1970, Aug. 2019, doi: 10.1109/TMI.2019.2898414.

[18] W. Lu, Y. Tong, Y. Yu, Y. Xing, C. Chen, and Y. Shen, “Deep Learning-Based Automated Classification of Multi-Categorical Abnormalities From Optical Coherence Tomography Images,” *Transl. Vis. Sci. Technol.*, vol. 7, no. 6, pp. 41–41, Nov. 2018, doi: 10.1167/TVST.7.6.41.

[19] M. Adhi and J. S. Duker, “Optical coherence tomography – current and future applications,” *Curr. Opin. Ophthalmol.*, vol. 24, no. 3, p. 213, May 2013, doi: 10.1097/ICU.0B013E32835F8BF8.

[20] A. E. Heidari *et al.*, “Optical Coherence Tomography as an Oral Cancer Screening Adjunct in a Low Resource Settings,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 25, no. 1, 2018, doi: 10.1109/JSTQE.2018.2869643.

[21] B. L. James *et al.*, “Validation of a point-of-care optical coherence tomography device with machine learning algorithm for detection of oral potentially malignant and malignant lesions,” *Cancers (Basel).*, vol. 13, no. 14, Jul. 2021, doi: 10.3390/CANCERS13143583/S1.

[22] Z. Yang, H. Pan, J. Shang, J. Zhang, and Y. Liang, “Deep-Learning-Based Automated Identification and Visualization of Oral Cancer in Optical Coherence Tomography Images,” *Biomedicines*, vol. 11, no. 3, p. 802, 2023, doi: 10.3390/biomedicines11030802.

[23] R. N. Goldan *et al.*, “Automated segmentation of oral mucosa from wide-field OCT images (Conference Presentation),” in *Proc.SPIE*, Apr. 2016, vol. 9698, p. 96980R, doi: 10.1117/12.2211122.

[24] C. Hill, “A Deep Learning Approach to Automated Segmentation of Oral Mucosa From OCT Images,” Simon Fraser University, Burnaby, 2020.

[25] C. Hill, C. Poh, C. MacAulay, and P. Lane, “Automated detection of the epithelial-stromal boundary in oral OCT images using deep learning,” in *Proc.SPIE*, Mar. 2022, vol. PC11952, p. PC119520G, doi: 10.1117/12.2610365.

[26] C. Hill, C. Poh, C. MacAulay, and P. Lane, “Epithelial segmentation of oral OCT with deep learning to quantify thickness and degree of stratification,” *https://doi.org/10.1117/12.2650484*, vol. 12354, p. 123540C, Mar. 2023, doi: 10.1117/12.2650484.

[27] T. Winslow, “Anatomy of the Oral Cavity,” *Medical and Scientific Illustration*, 2012. https://www.teresewinslow.com/head/wjwd7k37kdtlry87aqz7r13majf813 (accessed Feb. 08, 2023).

[28] M. Brizuela and R. Winters, “Histology, Oral Mucosa,” *StatPearls*, May 2023, Accessed: Jun. 05, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK572115/.

[29] C. Squier and K. A. Brogden, “The Organization of Oral Mucosa,” in *Human Oral Mucosa*, John Wiley & Sons, Ltd, 2011, pp. 9–17.

[30] C. Squier and K. A. Brogden, “Regional Differences in the Oral Mucosa,” in *Human Oral Mucosa*, John Wiley & Sons, Ltd, 2011, pp. 77–98.

[31] Stasio, Lauritano, Iquebal, Romano, Gentile, and Lucchese, “Measurement of Oral Epithelial Thickness by Optical Coherence Tomography,” *Diagnostics*, vol. 9, no. 3, p. 90, Aug. 2019, doi: 10.3390/diagnostics9030090.

[32] M. Albrecht, C. Schnabel, J. Mueller, J. Golde, E. Koch, and J. Walther, “In Vivo Endoscopic Optical Coherence Tomography of the Healthy Human Oral Mucosa: Qualitative and Quantitative Image Analysis,” *Diagnostics 2020, Vol. 10, Page 827*, vol. 10, no. 10, p. 827, Oct. 2020, doi: 10.3390/DIAGNOSTICS10100827.

[33] A. J. P. Klein-Szanto and H. E. Schroeder, “Architecture and density of the connective tissue papillae of the human oral mucosa,” *J. Anat*, vol. 123, no. 1, pp. 93–109, 1977.

[34] A. H. Fischer, K. A. Jacobson, J. Rose, and R. Zeller, “Hematoxylin and eosin staining of tissue and cell sections,” *CSH Protoc.*, vol. 2008, no. 5, May 2008, doi: 10.1101/PDB.PROT4986.

[35] D. R. Brenner *et al.*, “Projected estimates of cancer in Canada in 2022,” *CMAJ*,

vol. 194, no. 17, pp. E601–E607, May 2022, doi: 10.1503/CMAJ.212097.

[36] J. Seoane Lestón and P. Diz Dios, “Diagnostic clinical aids in oral cancer,” *Oral Oncol.*, vol. 46, no. 6, pp. 418–422, 2010, doi: 10.1016/j.oraloncology.2010.03.006.

[37] J. J. Pindborg, P. A. Reichart, C. J. C. Smith, I. van der Waal, and I. Van der Waal, *Histological typing of cancer and precancer of the oral mucosa*, 2nd ed. Springer Berlin Heidelberg, 1997.

[38] G. J. Kelloff and C. C. Sigman, “Assessing intraepithelial neoplasia and drug safety in cancer-preventive drug development,” *Nat. Rev. Cancer 2007 77*, vol. 7, no. 7, pp. 508–518, Jul. 2007, doi: 10.1038/nrc2154.

[39] I. van der Waal, “Potentially malignant disorders of the oral and oropharyngeal mucosa; terminology, classification and present concepts of management,” *Oral Oncol.*, vol. 45, no. 4–5, pp. 317–323, Apr. 2009, doi: 10.1016/J.ORALONCOLOGY.2008.05.016.

[40] L. Collins and S. Thavaraj, “Histological Aspects of Oral Potentially Malignant Disorders,” in *Oral Potentially Malignant Disorders: Healthcare Professional Training*, 1st ed., R. Albuquerque, V. Brailo, B. Carey, M. Diniz-Freitas, J.-C. Fricain, G. Lodi, L. Monteiro, and S. Ariyaratnam, Eds. London, 2022, pp. 29–34.

[41] H. Mortazavi, M. Baharvand, and M. Mehdipour, “Oral Potentially Malignant Disorders: An Overview of More than 20 Entities,” *J. Dent. Res. Dent. Clin. Dent. Prospects*, vol. 8, no. 1, p. 6, 2014, doi: 10.5681/JODDD.2014.002.

[42] K. Ranganathan and L. Kavitha, “Oral epithelial dysplasia: Classifications and clinical relevance in risk assessment of oral potentially malignant disorders,” *J. Oral Maxillofac. Pathol.*, vol. 23, no. 1, p. 19, Jan. 2019, doi: 10.4103/JOMFP.JOMFP_13_19.

[43] C. F. Poh, S. Ng, K. W. Berean, P. M. Williams, M. P. Rosin, and L. Zhang, “Biopsy and histopathologic diagnosis of oral premalignant and malignant lesions.,” *J. Can. Dent. Assoc.*, vol. 74, no. 3, pp. 283–288, Apr. 2008.

[44] B. W. Neville and T. A. Day, “Oral Cancer and Precancerous Lesions,” *CA. Cancer J. Clin.*, vol. 52, no. 4, pp. 195–215, 2002, doi: https://doi.org/10.3322/canjclin.52.4.195.

[45] W. Fujimoto J.and Drexler, “Introduction to Optical Coherence Tomography,” in *Optical Coherence Tomography: Technology and Applications*, J. G. Drexler Wolfgangand Fujimoto, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–45.

[46] W. Drexler, J. G. Fujimoto, and Special Section Guest Editors, “Optical Coherence Tomography in Ophthalmology,” *J. Biomed. Opt.*, vol. 12, no. 4, p. 041201, 2007, doi: 10.1117/1.2773734.

[47] S. Chen *et al.*, “Ultrahigh Resolution OCT Markers of Normal Aging and Early Age-related Macular Degeneration,” *Ophthalmol. Sci.*, vol. 3, no. 3, p. 100277, Sep. 2023, doi: 10.1016/J.XOPS.2023.100277.

[48] J. Olsen, J. Holmes, and G. B. E. Jemec, “Advances in optical coherence tomography in dermatology—a review,” *https://doi.org/10.1117/1.JBO.23.4.040901*, vol. 23, no. 4, p. 040901, Apr. 2018, doi: 10.1117/1.JBO.23.4.040901.

[49] R. Steiner, K. Kunzi-Rapp, and K. Scharffetter-Kochanek, “Optical Coherence Tomography: Clinical Applications in Dermatology,” *Med. Laser Appl.*, vol. 18, no. 3, pp. 249–259, Jan. 2003, doi: 10.1078/1615-1615-00107.

[50] I. K. Jang *et al.*, “Visualization of coronary atherosclerotic plaques in patients using optical coherence tomography: Comparison with intravascular ultrasound,” *J. Am. Coll. Cardiol.*, vol. 39, no. 4, pp. 604–609, Feb. 2002, doi: 10.1016/S0735-1097(01)01799-5.

[51] C. M. Peters *et al.*, “Fiber optic endoscopic optical coherence tomography (OCT) to assess human airways: The relationship between anatomy and physiological function during dynamic exercise,” *Physiol. Rep.*, vol. 9, no. 1, Jan. 2021, doi: 10.14814/PHY2.14657.

[52] L. P. Hariri *et al.*, “Laparoscopic optical coherence tomography imaging of human

ovarian cancer," *Gynecol. Oncol.*, vol. 114, no. 2, pp. 188–194, Aug. 2009, doi: 10.1016/J.YGYNO.2009.05.014.

[53] L. B. DaSilva, B. W. Colston, U. S. Sathyam, M. J. Everett, P. Stroeve, and L. L. Otis, "Dental OCT," *Opt. Express, Vol. 3, Issue 6, pp. 230-238*, vol. 3, no. 6, pp. 230–238, Sep. 1998, doi: 10.1364/OE.3.000230.

[54] J. G. Fujimoto, C. Pitris, S. A. Boppart, and M. E. Brezinski, "Optical coherence tomography: An emerging technology for biomedical imaging and optical biopsy," *Neoplasia*. 2000, doi: 10.1038/sj.neo.7900071.

[55] W. Jung and S. A. Boppart, "Optical coherence tomography for rapid tissue screening and directed histological sectioning," *Anal. Cell. Pathol.*, vol. 35, no. 3, pp. 129–143, 2012, doi: 10.3233/ACP-2011-0047.

[56] A. S. Pejcic, V. D. Zivkovic, V. R. Bajagic, and D. S. Mirkovic, "Histological changes of gingival epithelium in smokers and non-smokers," *Cent. Eur. J. Med.*, vol. 7, no. 6, pp. 756–760, 2012, doi: 10.2478/s11536-012-0050-8.

[57] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, no. 1, pp. 1–9, Oct. 2019, doi: 10.1186/S12916-019-1426-2/PEER-REVIEW.

[58] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2015.

[59] L. R. Dice, "Measures of the Amount of Ecologic Association Between Species," *Ecology*, vol. 26, no. 3, pp. 297–302, Jun. 1945, doi: 10.2307/1932409.

[60] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, 1950, doi: 10.1002/1097-0142.

[61] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLoS One*, vol. 10, no. 3, p. e0118432, Mar. 2015, doi: 10.1371/JOURNAL.PONE.0118432.

[62] A. Reinke *et al.*, "Common Limitations of Image Processing Metrics: A Picture Story," Apr. 2021, doi: 10.48550/arxiv.2104.05642.

[63] H. C. W. de Vet, L. B. Mokkink, C. B. Terwee, O. S. Hoekstra, and D. L. Knol, “Clinicians are right not to like Cohen’s κ,” *BMJ*, vol. 346, 2013, doi: 10.1136/BMJ.F2125.

[64] “Adjust image intensity values or colormap - MATLAB imadjust.” https://www.mathworks.com/help/images/ref/imadjust.html (accessed Mar. 20, 2023).

[65] “Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) - MATLAB pchip.” https://www.mathworks.com/help/matlab/ref/pchip.html (accessed Apr. 11, 2023).

[66] S. K. Warfield, K. H. Zou, and W. M. Wells, “Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation,” *IEEE Trans. Med. Imaging*, vol. 23, no. 7, p. 903, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[67] “BCEWithLogitsLoss — PyTorch 2.0 documentation.” https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html (accessed Apr. 20, 2023).

[68] “Z-Score Normalization: Definition & Examples - Statology.” https://www.statology.org/z-score-normalization/ (accessed Jul. 02, 2023).

[69] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015, doi: 10.1007/978-3-319-24574-4_28.

[70] C. Rivera, “Essentials of oral cancer,” *Int. J. Clin. Exp. Pathol.*, vol. 8, no. 9, p. 11884, 2015, Accessed: Jul. 02, 2023. [Online]. Available: /pmc/articles/PMC4637760/.

[71] A. M. D. Lee *et al.*, “Validation of Airway Wall Measurements by Optical Coherence Tomography in Porcine Airways,” *PLoS One*, vol. 9, no. 6, p. e100145, Jun. 2014, doi: 10.1371/JOURNAL.PONE.0100145.

[72] Y. Chen *et al.*, “Validation of human small airway measurements using endobronchial optical coherence tomography,” *Respir. Med.*, vol. 109, no. 11, pp. 1446–1453, Nov. 2015, doi: 10.1016/J.RMED.2015.09.006.

[73] A. F. Zuluaga, M. Follen, I. Boiko, A. Malpica, and R. Richards-Kortum, “Optical coherence tomography: A pilot study of a new imaging technique for noninvasive examination of cervical tissue,” *Am. J. Obstet. Gynecol.*, vol. 193, no. 1, pp. 83–88, Jul. 2005, doi: 10.1016/J.AJOG.2004.11.054.

[74] W.-J. Madore *et al.*, “Morphologic three-dimensional scanning of fallopian tubes to assist ovarian cancer diagnosis,” *https://doi.org/10.1117/1.JBO.22.7.076012*, vol. 22, no. 7, p. 076012, Jul. 2017, doi: 10.1117/1.JBO.22.7.076012.

Appendix A. Rater Annotation Training Slices

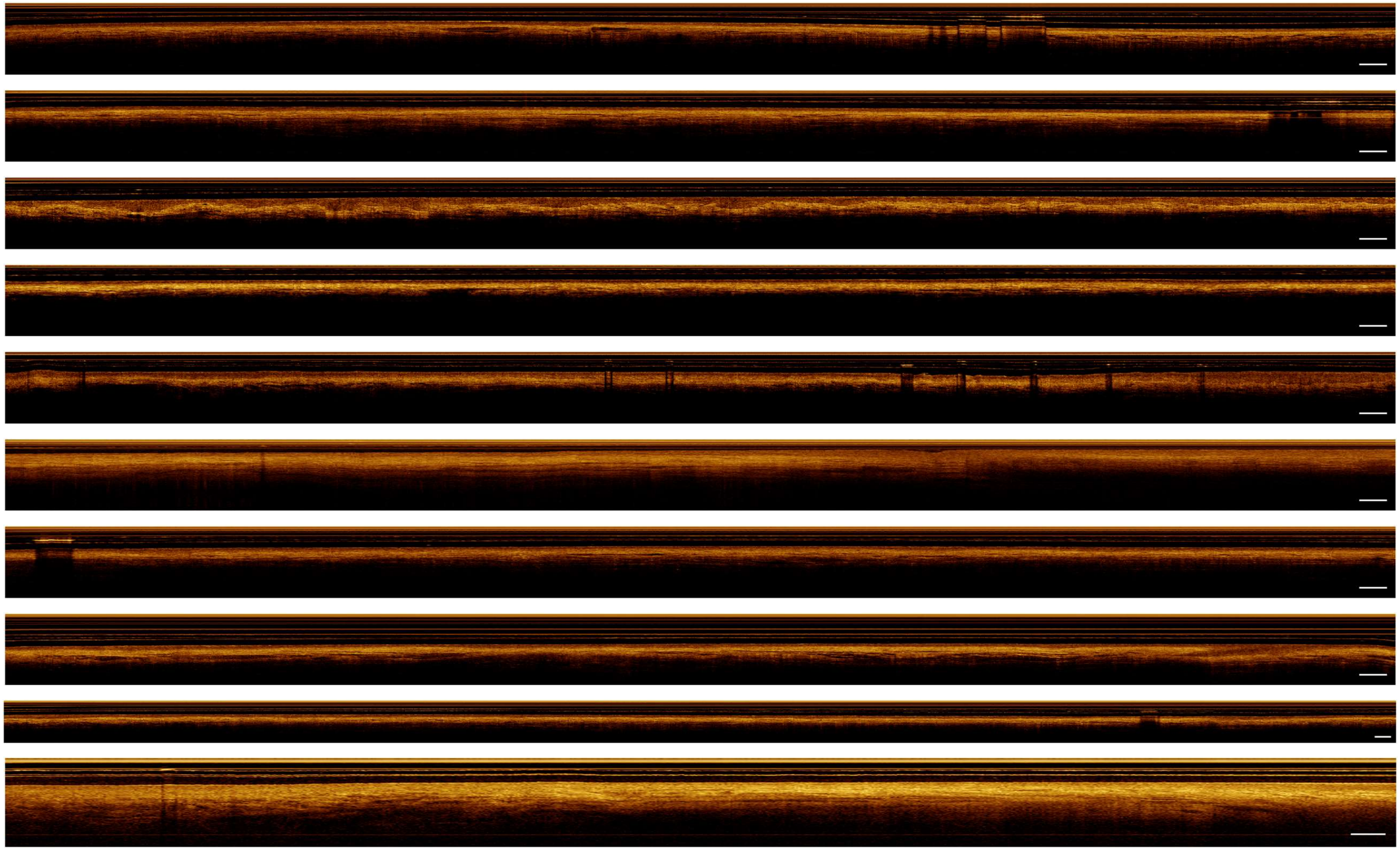


Figure A1. Longitudinal slices distributed for rater training

Appendix B. Network Parameters Details

Table B1. FOV Classification CNN Parameters

Layer		Values
Convolutional Layers		
x2	Convolution	kernel = (5,3); stride = (1,1); padding = (2,1)
	ReLu	In-place = true
	Pool	kernel =(2,2); stride = (1,1)
Classification Layers		
	Flatten	
	Linear	
	Relu	In-place = false
	Linear	
	Sigmoid	

Table B2. Epithelial surface segmentation u-net Parameters.

Layer		Values
Encoder		
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Pool	kernel = (2,2), stride = (2,2)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Pool	kernel = (2,2), stride = (2,2)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place
	Pool	kernel = (2,2), stride = (2,2)
Bottleneck		
x1	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
Decoder		
	Transposed Convolution	kernel = (2,2); stride = (2, 2); padding = (0,0)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Transposed Convolution	kernel = (2,2); stride = (2, 2); padding = (0,0)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)

	Batch Normalization	
	ReLu	In-place = true
	Transposed Convolution	kernel = (2,2); stride = (2, 2); padding = (0,0)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Convolution	kernel = (1,1); stride = (1,1); padding = (0,0)
	Sigmoid	

Table B3. Basement Membrane u-net Parameters.

Layer		Values
Encoder		
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Pool	kernel = (2,2), stride = (2,2)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Pool	kernel = (2,2), stride = (2,2)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Pool	kernel = (2,2), stride = (2,2)
Bottleneck		
x1	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
Decoder		
	Transposed Convolution	kernel = (2,2); stride = (2, 2); padding = (0,0)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Transposed Convolution	kernel = (2,2); stride = (2, 2); padding = (0,0)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true
	Transposed Convolution	kernel = (2,2); stride = (2, 2); padding = (0,0)
x2	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	Batch Normalization	
	ReLu	In-place = true

	Convolution	kernel = (1,1); stride = (1,1); padding = (0,0)
	Sigmoid	

Table B4. Artifact classification CNN Parameters

Layer		Values
Convolutional Layers		
x4	Convolution	kernel = (3,3); stride = (1,1); padding = (1,1)
	ReLu	In-place = true
	Pool	kernel =(2,2); stride = (1,1)
Classification Layers		
	Flatten	
	Linear	
	Relu	In-place = false
	Linear	
	Sigmoid	

Appendix C. Summary of Site and Diagnosis of Segmentation Networks Test Set

Table C1. Epithelial

Anatomical site	Diagnosis	No. Pullbacks
Buccal Mucosa	Dysplasia Grade 1	1
Gingiva	CIS	1
Lip	Benign	1
Ventral Tongue	Contralateral	9
	Dysplasia Grade 1	1
	Dysplasia Grade 2	2
	Dysplasia Grade 3	1
	OSCC	2
	Unknown Lesion	4
Lateral Tongue	Contralateral	6
	Hyperplasia	2
	Dysplasia Grade 3	2

Appendix D. Expanded Epithelial Surface Masks

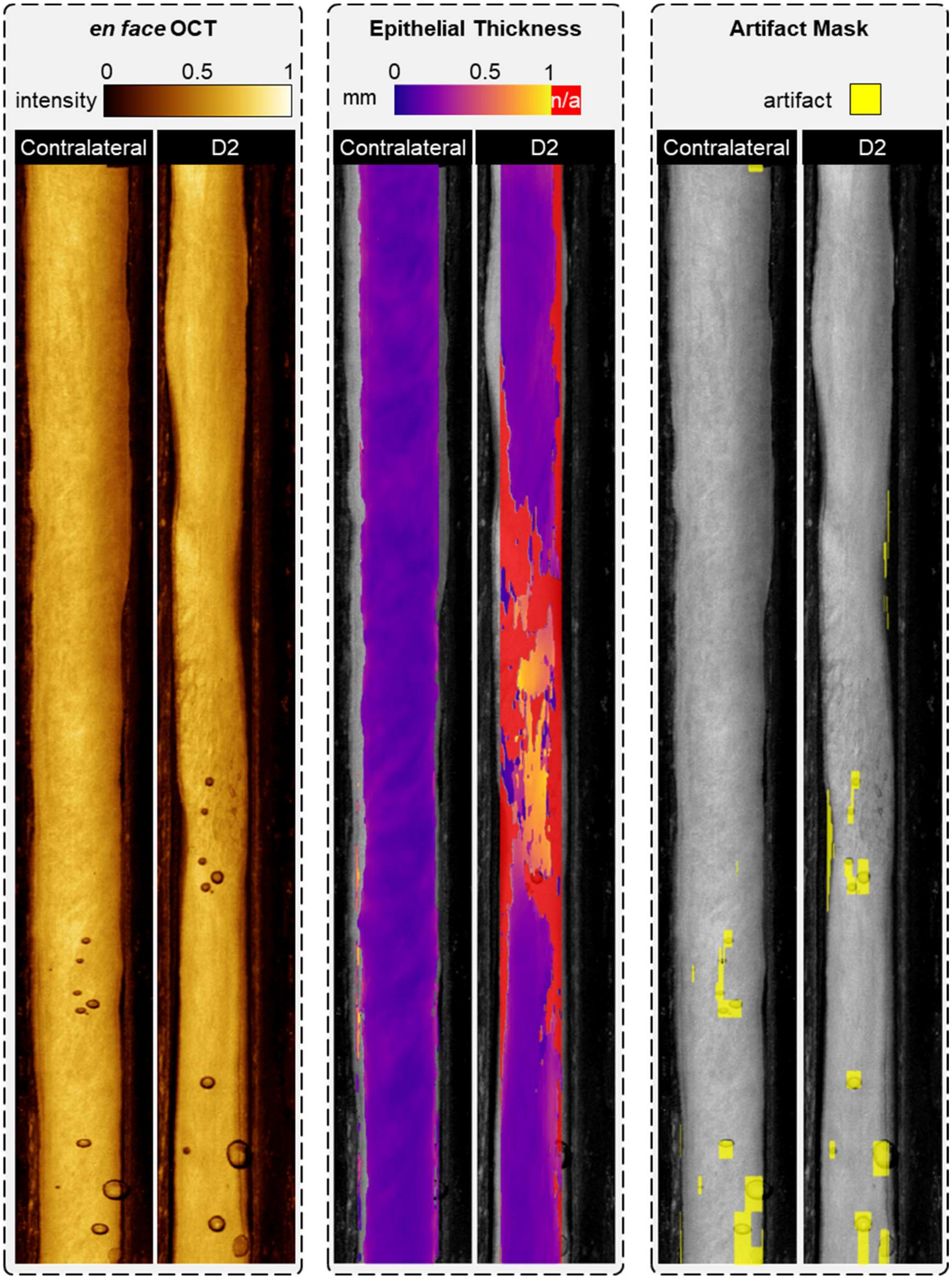


Figure D1. OCT *en face* of contralateral and dysplasia grade 2 (D2) lateral tongue; unannotated; with superimposed epithelial thickness map; with superimposed artifact locale mask.